Étude de différentes méthodes d'apprentissage supervisé pour le développement de tests diagnostiques basés sur des données métabolomiques

David Chardin

▶ **To cite this version:**

David Chardin. Étude de différentes méthodes d'apprentissage supervisé pour le développement de tests diagnostiques basés sur des données métabolomiques. Médecine humaine et pathologie. Université Côte d'Azur, 2023. Français. NNT : 2023COAZ6004 . tel-04143914

HAL Id: tel-04143914
https://theses.hal.science/tel-04143914v1

Submitted on 28 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES DE LA VIE ET DE LA SANTÉ

THÈSE DE DOCTORAT

Étude de différentes méthodes d'apprentissage supervisé pour le développement de tests diagnostiques basés sur des données métabolomiques

David CHARDIN

Laboratoire TIRO-MATOs UMR E4320

Présentée en vue de l'obtention du grade de docteur en Science de la Vie et de la Santé
d'Université Côte d'Azur
Dirigée par : Olivier Hubert

Soutenue le : 21 Mars 2023

Devant le jury, composé de :
Fanny Burel Vandenbos, PU-PH, CHU de Nice, Université Cote d'Azur
Marc Chadeau-Hyam, Professeur, Imperial College London
Olivier Humbert, PU-PH, Centre Antoine Lacassagne, Université Cote d'Azur
Laurent Suissa, PU-PH, CHU de Marseille, Aix-Marseille Université

ETUDE DE DIFFERENTES METHODES D'APPRENTISSAGE SUPERVISE POUR LE DEVELOPPEMENT DE TESTS DIAGNOSTIQUES BASES SUR DES DONNEES METABOLOMIQUES

Jury:

Rapporteurs

Pr. Marc Chadeau-Hyam, PHD, Professor in Computational Epidemiology and Biostatistics
Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London
St Mary's Hospital, Norfolk Place, W21PG, London

Pr. Laurent Suissa, MD-PHD, PU-PH,
Unité Neurovasculaire, Hôpital La Timone, CHU de Marseille (AP-HM)
Centre de recherche en CardioVasculaire et Nutrition (C2VN), Faculté de Médecine de Marseille, Aix-Marseille Université

Examinatrice

Pr. Fanny Burel Vandenbos, MD-PHD, PU-PH,
Laboratoire Central d'Anatomie Pathologique, CHU de Nice
UMR CNRS 7277-UMR INSERM 1091, Institut de Biologie Valrose, Université Côte d'Azur

Directeur de Thèse

Pr. Olivier Humbert, MD-PHD,
Service de Médecine Nucléaire du Centre Antoine Lacassagne,
Laboratoire TIRO-MATOs, UMR E4320, Université Côte d'Azur

Étude de différentes méthodes d'apprentissage supervisé pour le développement de tests diagnostiques bases sur des données métabolomiques

---

Résumé :

La métabolomique est une approche portant sur l'étude des petites molécules ou « métabolites » présents dans divers échantillons biologiques. Les différents domaines omiques : génomique, transcriptomique, protéomique et métabolomique, forment une chaine où chaque maillon va influencer les autres et pourra être influencé par des phénomènes externes. La métabolomique représente le dernier maillon de cette chaine, résultat de facteurs génétiques, pathologiques, environnementaux et toxicologiques et est ainsi le domaine omique qui se rapproche le plus du phénotype biologique.

Les analyses de métabolomique étant relativement peu couteuses et rapides, elles pourraient être utilisées en médecine, notamment pour élaborer de nouveaux tests diagnostiques.

Les données de métabolomique comportent un grand nombre de variables. Différentes méthodes de machine learning sont utilisées pour l'analyse statistiques de ces données de grande dimension. La méthode la plus utilisées est la méthode PLS-DA (Partial Least Squares Discriminant Analysis). Cependant, cette méthode présente certaines limites, notamment un risque de fausses découvertes lié à un sur-ajustement.

Dans le cadre de cette thèse, nous avons évalué de nouvelles méthodes de classification supervisée pour des applications cliniques de la métabolomique, notamment pour le développement de tests diagnostiques.

Nous présentons tout d'abord deux nouvelles méthodes de classification supervisée développées en collaboration entre biologistes, médecins et mathématiciens pour une utilisation en métabolomique : la méthode PD-CR (Primal Dual for Classification with Rejection) et un autoencodeur supervisé. Nous comparons ces méthodes à des méthodes couramment utilisées en métabolomique : PLS-DA, Standard Vector Machines (SVM), Random Forests et un réseau de neurone. Nous montrons ainsi que les nouvelles méthodes développées présentent des performances équivalentes ou supérieures aux méthodes courantes tout en sélectionnant des métabolites pertinents, dont le poids dans la classification est donné de manière facilement interprétable. Par ailleurs, ces méthodes incluent un score de probabilité pour chaque prédiction, qui nous semble particulièrement pertinent pour une utilisation dans un contexte médical.

Ensuite nous présentons les résultats d'une étude de métabolomique concernant des échantillons de tumeurs gliales congelés et fixés en paraffines. A l'aide d'une méthode de régression avec pénalisation L1 associée à un bootstrap nous avons développé deux modèles

permettant de classifier les tumeurs gliales selon leur statut mutationnel IDH et selon leur grade à partir de données de métabolomique obtenues sur échantillons congelés. Ces modèles étaient basés sur trois métabolites d'intérêt : le 2-hydroxyglutarate, l'amino-adipate et le guanidine-acetate. Nous avons ensuite montré que ces modèles pouvaient être appliqués sur des données de métabolomique obtenues sur des échantillons fixés en paraffine avec des performances correctes : prédiction du statut mutationnel IDH avec une sensibilité 70.6% de et une spécificité de 80.4% et prédiction du grade avec une sensibilité de 75% et une spécificité de 74.5%. Nous avons ainsi montré qu'il était possible de réaliser des analyses de métabolomique sur échantillons fixés en paraffine et d'en tirer des résultats pertinents.

L'analyse ciblée de nouveaux échantillons permettrait de valider ces modèles et de les utiliser en pratique courante en complément des techniques déjà disponibles. De plus, l'exploration des phénomènes biologiques à l'origine de l'association entre le grade de malignité des tumeurs gliales et l'aminoadipate et le guanidinoactetate pourrait permettre de mieux comprendre leur cancérogénèse.

Mots clés : Métabolomique, Biomarqueurs, Apprentissage automatique.

Evaluation of different supervised learning methods for the creation of diagnostic tools based on metabolomic data.

Abstract :

Metabolomics is a recent field of research concerning the study of small molecules or « metabolites » in biological samples. The different omics fields: genomics, transcriptomics, proteomics and metabolomics, form a chain in which each link influences the others and is influenced by external factors. Metabolomics represent the last link of this chain, resulting of genetic, pathologic, environmental and toxicological factors, and is thus the omics field closest to the biological phenotype.

Since metabolomic studies are relatively fast and inexpensive, they could be used in routine medical practice, particularly for diagnostic testing.

Metabolomic data most frequently include high numbers of variables. Different machine learning methods are used for the statistical analysis of these high dimensional datasets. The most frequently used method is PLS-DA (Partial Least Squares Discriminant Analysis). However, this method has some drawbacks, including a risk of false discoveries due to overfitting.

In this work, we evaluated new supervised classification methods for clinical applications of metabolomics, particularly for diagnostic testing.

We first introduce two new classification methods, created through a collaboration between biologists, physicians and mathematicians: the PD-CR method (Primal Dual for Classification with Rejection) and a supervised autoencoder. We compare these methods to the most frequently used methods in this setting: PLS-DA, Standard Vector Machines, Random Forests and neural networks. Hence, we show that these new methods have similar or higher performances as the classical methods, while selecting biologically relevant metabolites for which the weights in the classification are given in a straightforward and easily interpretable manner. Moreover, these methods include a probability score for each prediction, which seems particularly relevant for medical applications.

We then report the results of a metabolomic study performed on frozen and formalin fixed glial tumor samples. Using an L1 penalized regression method associated with a bootstrap method we created two models to classify glial tumors according to their IDH mutational status and their grade. These models were trained on metabolomic data from frozen samples and lead to the selection of three metabolites: 2-hydroxyglutarate, aminoadipate and guanidinoacetate. When testing these models on metabolomic data obtained on fixed glial tumor samples, they

revealed good classification results: IDH mutational status prediction with a sensitivity of 70.6% and a specificity of 80.4% and grade prediction with a sensitivity of 75% and a specificity of 74.5%. Hence, we have shown that performing a metabolomic analysis on fixed samples is possible and can lead to promissing results.

Targeted analysis on new tumor samples could be performed to validate our models and lead to applications in routine practice, complementing pre-existing techniques. Moreover, exploring the biological phenomenons underlying the association of glial tumor grade and aminoadipate and guanidinoacetate could lead to a better understanding of these tumors and their carcinogenesis.

Key words : Metabolomics, Biomarkers, Machine Learning.

Table des matières

I. INTRODUCTION A LA METABOLOMIQUE

A) Définition

La métabolomique est une approche portant sur l'étude des petites molécules (de taille inférieure à 1500 Da) ou « métabolites » présents dans divers échantillons biologiques. Cette approche introduite en 1998 [1] complète les approches « omiques » antérieures que sont la génomique, la transcriptomique et la protéomique.

Au sein d'un système biologique, les différents domaines omiques forment une chaine où chaque maillon va influencer les autres et pourra être influencé par des phénomènes externes. La métabolomique représente le dernier maillon de cette chaine, résultat de facteurs génétiques, pathologiques, environnementaux et toxicologiques (Figure 1) [2]. Ainsi la métabolomique est probablement le domaine omique qui se rapproche le plus du phénotype biologique et présente de ce fait un intérêt particulier dans divers domaines de recherche.

Par ailleurs, les analyses de métabolomique étant relativement peu couteuses et rapides, elles pourraient être réalisées en pratique courante. En effet, en l'appliquant sur des échantillons biologiques de patients, la métabolomique pourrait permettre de diagnostiquer certaines pathologies ou d'identifier certaines sous-classes de pathologies.

Comme les autres données omiques, les données de métabolomique comportent, en règle générale, un grand nombre de variables pour chaque échantillon analysé. Des méthodes statistiques adaptées à ces données de grandes dimensions doivent donc être utilisées. Les méthodes de machine learning, faisant actuellement l'objet de développements rapides, sont particulièrement adaptées à cette situation.

Dans le cadre de cette thèse, nous nous sommes particulièrement intéressés aux méthodes de machine learning utilisables pour les applications cliniques de la métabolomique, notamment pour le développement de tests diagnostiques.

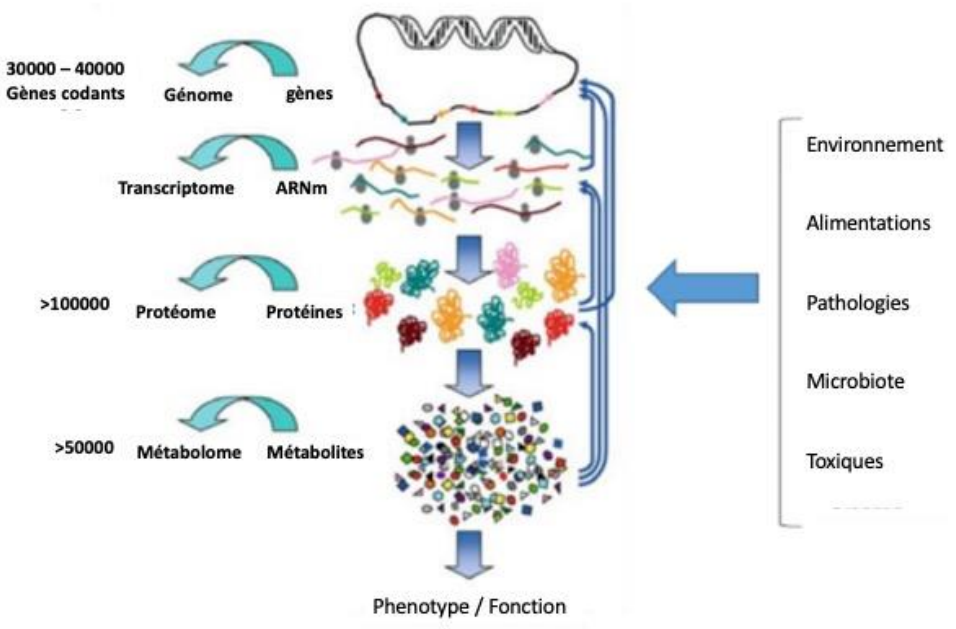

Figure 1. Interactions complexes, multi-directionnelles, entre les différents niveaux des systèmes biologiques.

B) Workflow général

Une étude de métabolomique s'organise en cinq étapes (figure 2).

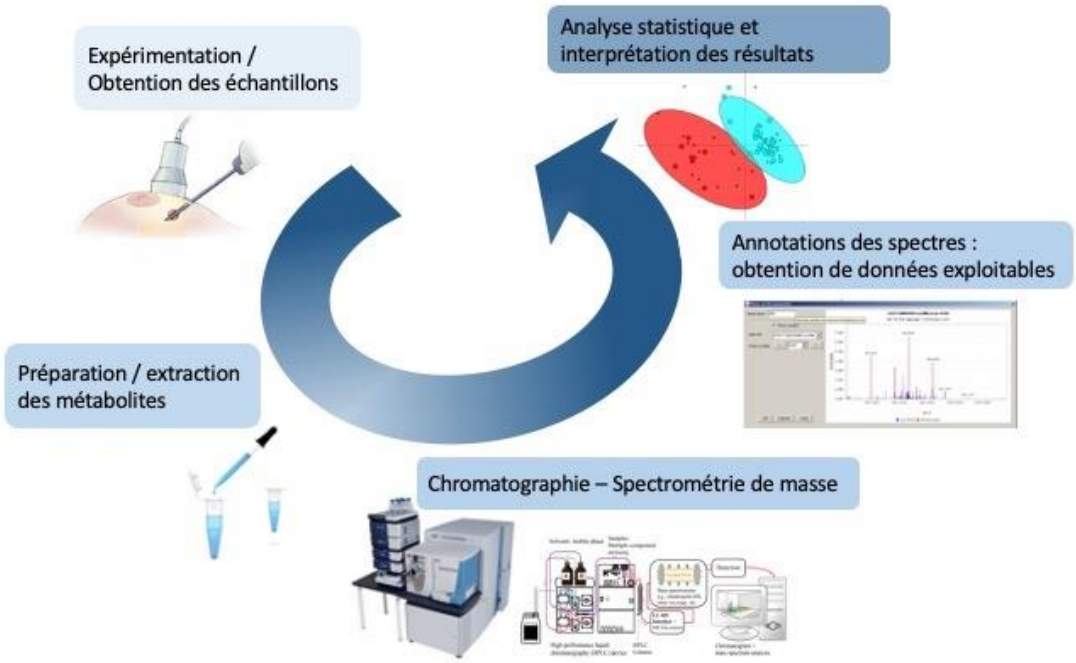

Figure 2. Workflow général d'une étude métabolomique

1. Obtention des échantillons

Il est possible d'obtenir des données de métabolomique pour de nombreux types d'échantillons biologiques (plantes, tissus, sang, urines etc.). Cependant, certains échantillons sont plus faciles à obtenir que d'autres. Par exemple, pour des applications médicales de routine, il sera plus aisé d'utiliser des prélèvements d'urine ou de sang que des prélèvements de tissus. En effet, les prélèvements de fluides sont beaucoup plus fréquents et sont moins invasifs que les prélèvements de tissus.

En revanche, l'analyse de fluides comme le sang ou les urines expose à un plus grand risque de biais, liés par exemple à la prise de médicaments, l'alimentation ou au rythme circadien [3,4] De plus, si un métabolite est produit en petite quantité au niveau d'un tissus pathologique, celui-ci pourrait n'être détectable qu'au sein de ce tissu.

Après le prélèvement, il sera en général nécessaire de réaliser une étape figeant les métabolites de l'échantillon. Cette étape fixe les différentes réactions biochimiques à un instant donné. Elle se fait le plus souvent par congélation de l'échantillon.

2. Préparation des échantillons puis extraction et purification des métabolites

Pour analyser les métabolites présents dans les échantillons, il faut les extraire, les purifier et les reprendre en solution. Ceci implique des étapes de broyage mécanique et de précipitation par solvants organiques des échantillons, suivi d'étapes de centrifugation puis récupération des surnageants. Les métabolites extraits vont dépendre du solvant utilisé. Un des solvant les plus utilisé est le méthanol. L'objectif de ces étapes est d'obtenir une solution contenant un maximum de métabolites d'intérêt et le moins possible de protéines ou déchets membranaires.

Dans le cadre de cette thèse, nous avons travaillé sur échantillons tumoraux congelés ou fixés en paraffine. Le détail de la préparation de ces échantillons est résumé dans l'article 3.

3. Spectrométrie de masse

La métabolomique a pu être développée grâce aux avancées technologiques récentes en spectrométrie de masse (MS) à haute résolution telles que les technologies d'analyseurs en temps de vol, orbitrap ou à résonance cyclotronique d'ion.
Les spectromètres de masse sont composés d'un système d'introduction de l'échantillon, d'une source d'ionisation, d'un analyseur en masse et d'un détecteur associé à un système de traitement.

En métabolomique, **le système d'introduction** est en général associé à une étape de chromatographie en phase gazeuse (GC) ou en phase liquide (LC). Cela permet d'augmenter le nombre de métabolites détectés, en réalisant une première étape de séparation des métabolites, basée sur leurs propriétés physico-chimiques. Cette première séparation permet D'optimiser la réponse en spectrométrie de masse.

La source d'ionisation sert à vaporiser les molécules et à les ioniser. Elle peut être utilisée soit en mode positif, pour étudier les ions positifs, soit en mode négatif, pour étudier les ions négatifs. Plusieurs types de sources existent et sont utilisées en fonction des molécules analysées. Les sources les plus fréquemment utilisées en combinaison avec la chromatographie en phase gazeuse sont l'ionisation électronique (EI) et l'ionisation chimique (CI). Les sources les plus fréquemment utilisées en combinaison avec la chromatographie en phase liquide sont l'éléctronébulisateur ou électrospray (ESI) et l'ionisation chimique à pression atmosphérique (APCI).

Une fois les molécules vaporisées et ionisées, celles-ci sont soumises à un courant électrique dans **l'analyseur**, et forment ainsi un courant ionique. Le **détecteur** permet ensuite de transformer ce courant ionique en courant électrique, dont le signal pourra être traité informatiquement pour obtenir des spectres de masses. Il existe différents types de couples **analyseurs-détecteurs**. Leur objectif est toujours de mesurer le rapport masse/charge des ions présents, ainsi que l'intensité du signal produit par ces ions.

L'analyseur en temps de vol mesure le temps nécessaire aux ions soumis à un courant de tension connue pour parcourir une distance connue pour en déduire leurs rapports m/z.
L'analyseur à résonance cyclotronique d'ion impose un mouvement cyclotronique aux ions. Les ions de même m/z sont mis en phase en leur appliquant une tension de fréquence

correspondant à leur fréquence cyclotronique. L'analyse des fréquences du courant ionique obtenu permet de déterminer le rapport m/z des ions par transformée de Fourier.
L'analyseur Orbitrap se compose d'une électrode creuse, à l'intérieur de laquelle est placée coaxialement une électrode en forme de fuseau. Il permet d'imposer un champ électrostatique quadro-logarithmique aux ions. Les ions de même m/z seront sur la même trajectoire circulaire qui oscille axialement autour de l'électrode centrale. Le courant induit par ces oscillations permet par une transformée de Fourier de déterminer les m/z.

La spectrométrie de masse en tandem (MS/MS) permet de mieux identifier les molécules analysées en les fragmentant. En effet, celle-ci combine une première analyse de spectrométrie de masse d'ions « parents » (MS1) et une seconde analyse de spectrométrie des ions « fils » obtenus après fragmentation de ces ions « parents » au sein d'une chambre de collision (MS2). Cette étape permet notamment de différencier des ions parents isomères, qui auront le même rapport m/z mais dont les fragments seront différents.

4. Annotation des spectres de masses

Les données de masses obtenues suite à l'étape précédente doivent être traitées informatiquement pour en extraire une information exploitable. L'objectif est d'individualiser les métabolites présents dans les échantillons et de les quantifier. Un des logiciels les plus utilisé pour réaliser ce traitement informatique est le logiciel MzMine [5].

Différentes étapes et techniques de traitement du signal sont utilisées. Celles-ci permettent d'individualiser les pics présents dans les données et de leur associer un rapport m/z ainsi qu'un temps de rétention.

Une des difficultés rencontrées en métabolomique est liée à l'hétérogénéité des données présentes dans les données de LC-MS. En effet, celles-ci incluent d'une part du bruit, pouvant masquer certains pics ou entrainer des identifications erronées de pics, et d'autre part de nombreux pics « perturbateurs », correspondant à des formes ioniques minoritaires de certains métabolites ou à des fragments de métabolites, pouvant également entrainer des identifications erronées. Différentes méthodes de filtrage existent pour éliminer ces pics perturbateurs [6,7]. Cependant, plus le filtrage utilisé est important et plus le risque d'éliminer de véritables métabolites d'intérêt est élevé.

En mesurant le rapport masse/charge (m/z) de chaque pic avec une grande précision (de l'ordre de 5 ppm avec la technologie orbitrap notamment), on peut identifier à quels ions correspondent chaque pic en confrontant les valeurs de m/z mesurées avec des bases de données. Dans un premier temps la valeur de m/z d'un ion « parent » (MS1) sera confrontée à une base de données de valeurs de m/z telle que HMDB (Human Metabolome Database)[8] (Figure 3A). Puis, pour confirmer l'identification des ions d'intérêt ou pour différencier plusieurs isomères potentiels, on confrontera leurs spectres MS2 à des bases de données de spectres MS2 telle que METLIN [9] (Figure 3B).

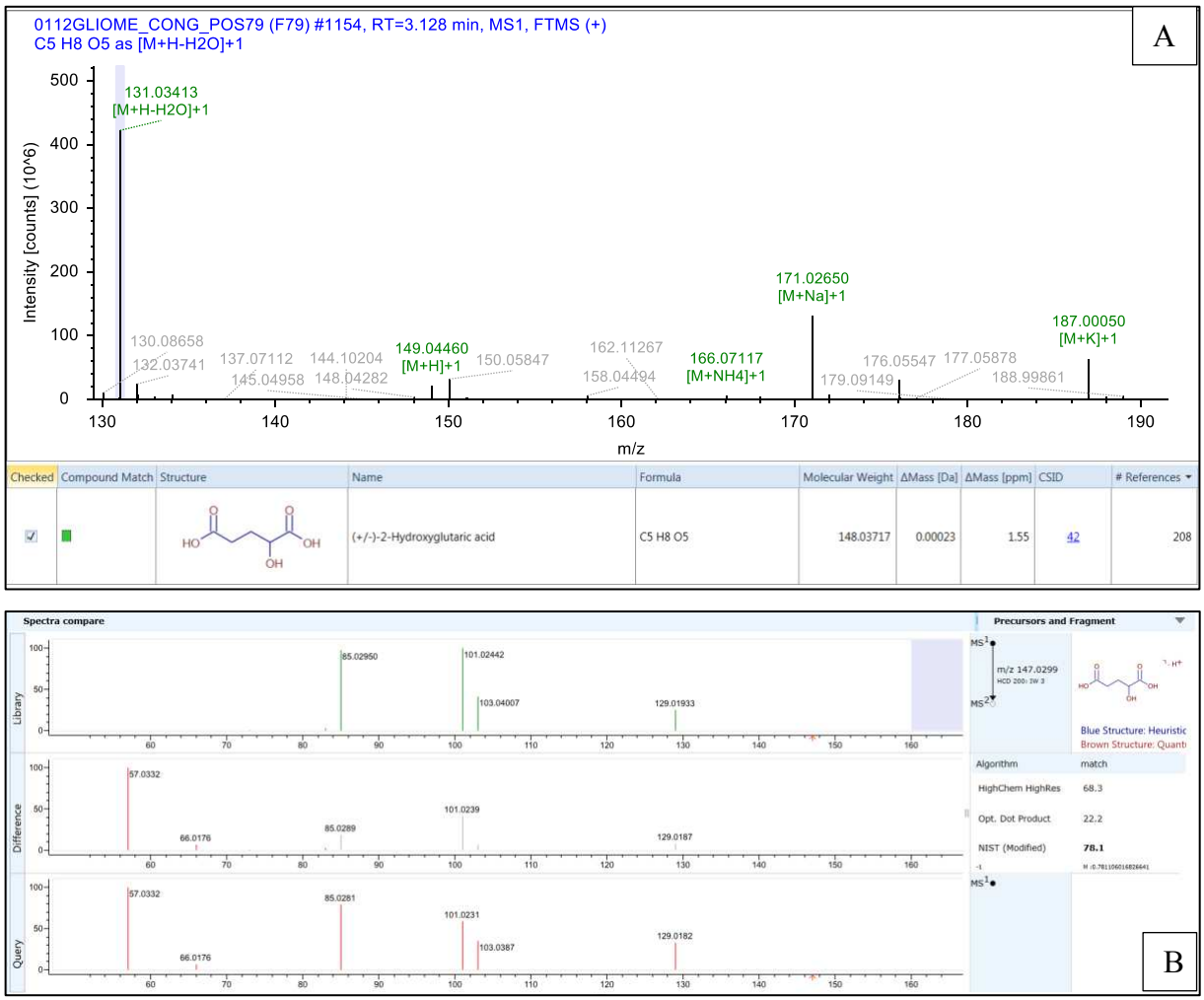

Figure 3. Exemple d'identification d'un métabolite : le 2-Hydroxyglutaric acid. A. Identification de différents pics correspondant à différentes formes ioniques du 2-Hydroxyglutaric acid. *On identifie également des isotopes de la forme majoritaire ([M+H-*

H_2O]+). La différence entre le m/z mesuré et le m/z de référence, enregistré dans la base de données, est de 1.55 ppm.

B. Confirmation de l'identification par confrontation du spectre MS2 mesuré au spectre MS2 de référence du 2-Hydroxyglutaric acid. *Le pourcentage de recouvrement des spectres est de 78.1%.*

5. Analyse statistique puis interprétation des résultats

Les données obtenues après ce traitement informatique correspondent à une liste de quelques centaines à milliers d'ions présents dans les échantillons, associés à leurs intensités, formant une matrice. L'analyse de ces nombreuses variables, dont plusieurs sont corrélées, nécessite l'utilisation de méthodes multivariées adaptées.

En effet, l'utilisation de tests univariés, comme le test de Student ou l'analyse de variance, pour évaluer la significativité d'éventuelles différences observées entre différentes classes d'échantillons, est limitée par le nombre de métabolites à analyser. En effet, la répétition de tests univariés pour chaque métabolite implique un risque rapidement croissant de faux positif [10,11]. Pour pallier ce risque, on peut appliquer des méthodes de correction de la valeur p (par exemple, la correction de Bonferroni [12]), mais ces méthodes majorent nettement la puissance statistique nécessaire pour mettre en évidence une différence significative.

Au contraire, les méthodes multivariées permettent d'analyser simultanément de nombreuses variables et sont donc les méthodes les plus utilisées en métabolomique. La différence entre ces méthodes « multivariées » et les méthodes de « machine learning », de popularité croissante, n'est pas claire et semble essentiellement se résumer à l'objectif final de chacune d'elles. Il apparait que les méthodes « multivariées » ont pour but d'expliquer le lien existant entre des variables explicatives et une observation en identifiant un modèle mathématique liant ces variables explicatives à l'observation, alors qu'en « machine learning » le but est de faire des prédictions à partir de ce type de modèle.

II. INTRODUCTION A L'APPRENTISSAGE AUTOMATISE OU « MACHINE LEARNING »

A) Définitions

L'intelligence artificielle (IA) désigne un ensemble de sciences, théories et techniques dont le but est la reproduction des capacités cognitives ou décisionnelles d'un être humain par une machine [13].

On distingue deux types d'IA. L'IA « forte », qui a pour ambition de développer des modèles permettant la création de machines dotées d'esprit et de conscience, et l'IA « faible », centrée sur une tâche donnée, entraînée pour résoudre un problème précis. Les techniques d'apprentissage automatique utilisées en biologie et en médecine appartiennent à l'IA « faible ».

L'apprentissage automatique ou « apprentissage machine » ou « machine learning » désigne un ensemble de techniques qui permettent de construire un modèle mathématique à partir de données, en incluant un grand nombre de variables, afin de réaliser une tâche. L'objectif est de créer un modèle (F) reliant des variables explicatives (X) à une ou plusieurs variables cibles (connues à l'avance ou non) (Y) par le biais de formules plus ou moins complexes, incluant un certain nombre de paramètres [14]. Les paramètres sont configurés au fur et à mesure lors d'une phase d'« apprentissage » sur un jeu de données d'entrainement. Une fois le modèle construit, il pourra être testé sur de nouveaux jeux de données de « test » pour s'assurer de sa capacité à généraliser son résultat sur de nouvelles données.

Les différentes méthodes d'apprentissage machine sont choisies en fonction de la nature des tâches à accomplir. Ces méthodes sont habituellement classées en 3 catégories : apprentissage supervisé, non supervisé et par renforcement (figure 4). Par ailleurs, depuis quelques années, une forme particulière d'apprentissage automatique, nommée « apprentissage profond » ou « deep learning » connait une popularité grandissante du fait de ses bonnes performances [14].

<u>Apprentissage supervisé</u> (données d'entrainement : X et Y connus)

L'apprentissage supervisé est utilisé pour créer des modèles prédictifs à partir de données incluant des variables explicatives et des variables cibles connues. Si la variable cible Y est une variable quantitative (ex : mesure ou comptage), on utilisera un modèle de régression, si celle-ci est qualitative (ex : catégorie), on utilisera un modèle de classification. C'est le type d'apprentissage le plus utilisé en biologie et en médecine. Par exemple, un modèle de classification supervisé peut être construit par apprentissage sur des données de métabolomique concernant des patients atteints d'une pathologie et des patients indemnes de cette même pathologie, pour prédire, à partir de données métabolomiques provenant d'un nouveau patient, si celui-ci est malade.

<u>Apprentissage non supervisé</u> (données d'entrainement X connu, mais Y inconnu)

L'apprentissage non supervisé est utilisé pour identifier des ensembles présentant des caractéristiques communes au sein d'un jeu de données. On peut utiliser des méthodes d'apprentissage non supervisé pour réduire la dimension de grands jeux de données, pour trouver des liens entre les données ou pour mettre en évidence des groupes similaires au sein d'une population (clustering). Ainsi, Perou et al ont utilisé ce type de méthode pour identifier 5 sous-types moléculaires de tumeurs du sein, à partir de données de génomique, présentant des caractéristiques biologiques et évolutives différentes [15]. En métabolomique, ce type de méthode a également permis d'identifier des sous-groupes de tumeurs du sein [16]. Ces méthodes permettent aussi de mettre en évidence des biais expérimentaux comme un effet batch, lié aux conditions de préparation des échantillons ou de paramétrage de l'analyse de LC-MS, ou des « outliers », échantillons aux caractéristiques extrêmes ou aberrantes.

<u>Apprentissage par renforcement</u> (Y est un objectif connu, X est découvert progressivement)

Ce type d'apprentissage est moins utilisé en biologie et en médecine. Il permet de résoudre des problèmes de manière séquentielle. Ce type d'apprentissage est notamment utilisé dans le cadre de jeux vidéo, par exemple pour gagner une partie d'échecs [17]. Les variables explicatives ne sont pas connues initialement mais sont « découvertes » de manière séquentielle, en enchainant les essais/erreurs. L'algorithme aura pour but de se rapprocher le plus possible d'un objectif représenté par une fonction de récompense en s'adaptant à chaque essais/erreurs. En médecine, ce type d'apprentissage pourrait être utilisé pour optimiser une séquence thérapeutique [18].

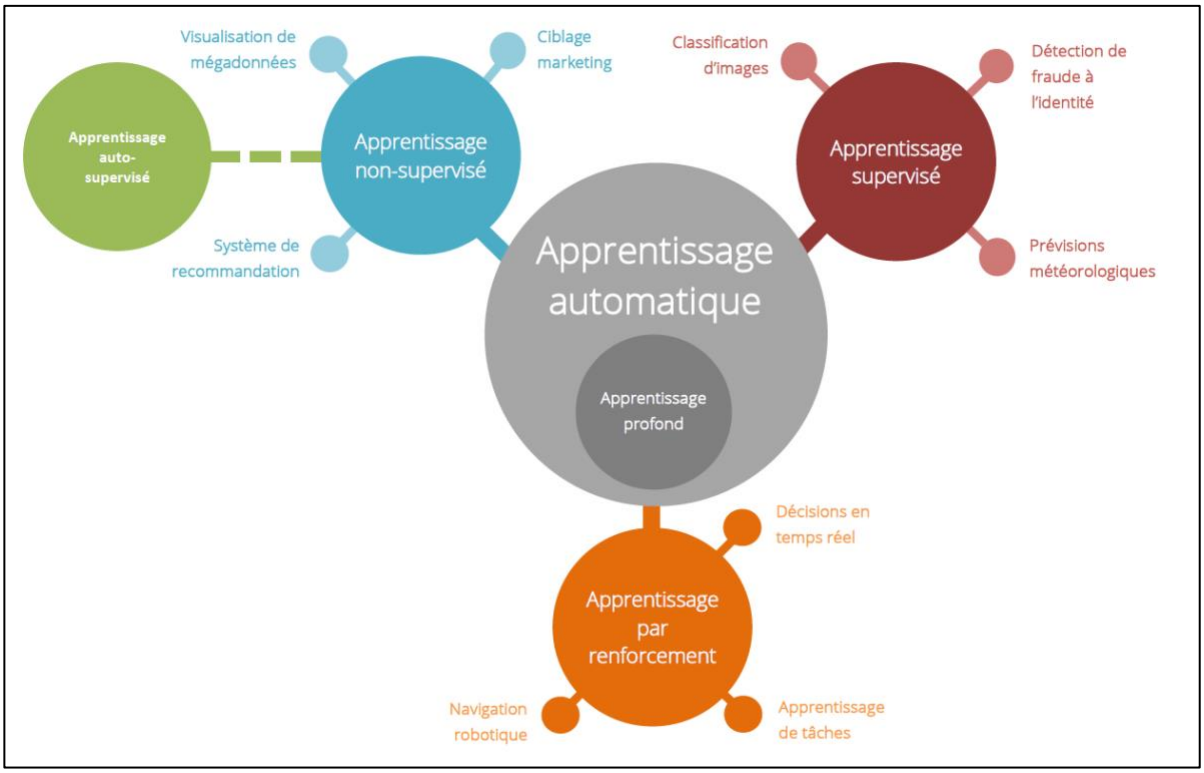

Figure 4. Représentation des différentes catégories d'apprentissage machine et exemples d'applications. *Figure issue du glossaire sur l'intelligence artificielle du conseil de l'Europe*

<u>Apprentissage profond ou « Deep learning »</u>

L'apprentissage profond ou « deep learning » regroupe un ensemble de techniques de machine learning basées sur des réseaux de neurones artificiels, incluant des couches « cachées » (plus de détails dans la section « Artificial Neural Networks » plus loin)

B) La malédiction de la dimension et ses solutions

La malédiction de la dimension (« curse of dimensionality ») est un terme employé par Richard Bellman dans les années 1950 [19] dans le domaine de l'optimisation dynamique, qui désigne divers phénomènes liés aux données de grandes dimensions. Ces phénomènes existent également en apprentissage automatique.

Intuitivement, s'il est possible d'inférer la distribution de probabilité d'une population dans un espace d'une seule dimension (une variable) à partir d'un nombre restreint d'observations,

beaucoup plus d'observations sont nécessaires pour inférer une distribution de probabilité dans un espace bidimensionnel ou tridimensionnel (figure 5). Dans le cadre de l'apprentissage automatique, l'espace comporte fréquemment quelques centaines, milliers ou dizaines de milliers de dimensions. La malédiction de la dimension correspond au fait que, lorsque le nombre de dimensions augmente, le volume de l'espace croît si rapidement que les données deviennent « isolées » ou « éparses » (« sparse » en anglais) ce qui limite les possibilités d'en tirer des inférences.

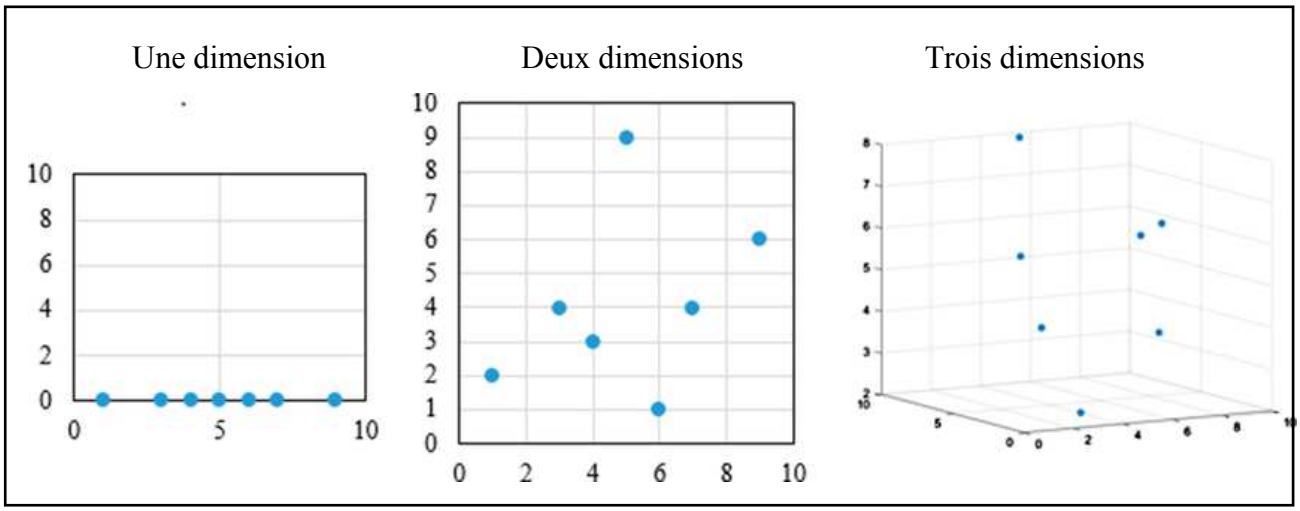

Figure 5. Illustration de la malédiction de la dimension

Parallèlement, la malédiction de la dimension peut également désigner le fait qu'il sera plus facile d'entrainer un modèle avec une base de données ne comportant que quelques variables discriminantes indépendantes qu'avec une base de données comportant quelques variables discriminantes mélangées au sein de nombreuses variables redondantes ou non discriminantes.

Une des solutions au problème de la malédiction de la dimension consiste à réduire le nombre de dimensions en réduisant le nombre de variables explicatives. Deux méthodes existent : l'une consiste à sélectionner un nombre restreint de variables représentatives, l'autre consiste à représenter l'ensemble des variables explicatives par un petit nombre de variables latentes qui « résument » l'information contenue dans plusieurs variables.

Les méthodes de sélection de variables permettent de diminuer le nombre de variables non discriminantes et redondantes. Le risque étant d'éliminer des variables qui n'apportent que peu d'informations isolément mais apportent une information pertinente une fois combinées entre

elles. La sélection de variables peut se faire par des méthodes de filtrage basées sur des calculs des performances individuelles de chaque variable ou par des méthodes de régularisation ou pénalisation (voir section III. F).

Les méthodes utilisant des variables latentes permettent de regrouper l'ensemble de l'information apportée par un grand nombre de variables dans une seule « super-variable ». Cela permet notamment de conserver l'information apportée par une combinaison de variables. Cependant, si ces méthodes sont appliquées à des bases de données comportant un grand nombre de variables non-discriminantes, les variables latentes créées risquent de ne comporter que peu d'information utile. L'analyse en composantes principales est une des méthodes de réduction de dimension les plus utilisées (voir section III. A).

C) Évaluation des modèles en machine learning

Métriques

Pour être utile un modèle doit être performant et généralisable.
Les performances vont être évaluées via différentes métriques.

Pour un modèle de classification, les métriques utilisées sont basées sur le comptage des erreurs de classification, pouvant être résumé dans un tableau de contingence (figure 6). La métrique la plus utilisée est l'**exactitude** ou « accuracy » en anglais. Celle-ci correspond au nombre d'échantillons correctement classés, rapporté au nombre total d'échantillons. La **précision** (dénommée Valeur Prédictive Positive ou VPP dans le cadre de l'évaluation de tests diagnostiques en médecine), ainsi que le **rappel** (dénommée « sensibilité » dans le cadre de l'évaluation de tests diagnostiques) sont également utilisées.

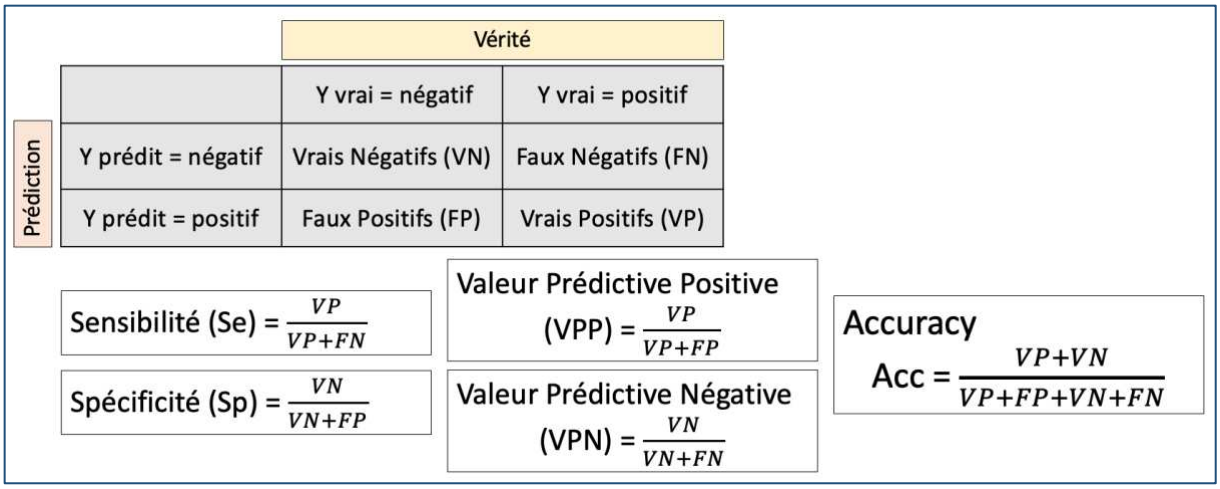

Figure 6. Présentation d'un tableau de contingence et de métriques associées. *Le cas présenté correspond à une classification binaire, cependant le tableau de contingence et les métriques associées peuvent être étendu pour de la classification multiclasse.*

N.B. Dans le cas où les classes sont non équilibrées (beaucoup de Y_{vrai} positifs mais très peu de Y_{vrai} négatifs), l'exactitude peut être artificiellement augmentée (Par exemple : s'il y a 85% de Y_{vrai} positifs et 15% de Y_{vrai} négatifs dans la cohorte, un modèle prédisant uniquement des résultats positifs aura tout de même une exactitude de 85%). On peut utiliser une exactitude équilibrée (« balanced accuracy »), pour limiter ce phénomène. Celle-ci prend en compte le nombre d'échantillon correctement classifié pour chaque classe.

Pour un modèle de régression, l'erreur est quantifiée en comparant les valeurs prédites aux valeurs attendues. La somme des carrés des résidus (**RSS** : Residual Sum of Squares) correspond à la mesure la plus simple de cette différence. Cependant celle-ci est d'autant plus grande qu'il n'y a de données testées. L'erreur quadratique moyenne (**MSE** : Mean Squared Error) est obtenue en normalisant la RSS par rapport au nombre de données testées et n'est ainsi pas affectée par ce phénomène. Cependant, la MSE ne permet pas de se rendre compte de l'importance de l'erreur sans connaitre la dimension attendue de la valeur à prédire (une erreur de dix centimètres est très faible pour un modèle concernant les mouvements planétaires mais considérable pour un modèle concernant la neuro-anatomie). L'erreur carrée relative (**RSE** : Relative Squared Error) est obtenue en normalisant la RSS par rapport à somme des écarts à la moyenne des valeurs tests. Le **coefficient de détermination** correspond au complémentaire à 1 de la RSE, il est également noté $\mathbf{R^2}$ car il correspond au carré du coefficient de corrélation de Pearson, noté R.

D) Notion d'underfitting et d'overfitting

Le caractère généralisable du modèle est évalué en testant ses performances sur différents jeux de données. En général, trois jeux de données sont utilisés : un jeu d'entrainement, un jeu de validation et un jeu de test. Le jeu d'entrainement et de validation sont en général obtenus en dichotomisant un même jeu de données.

L'apprentissage du modèle se fait sur les données d'« entrainement ». L'évaluation des performances du modèle sur ces données d'entrainement permet d'évaluer l'ajustement ou « fit » du modèle. Ensuite, le modèle est testé sur des données de « validation est de « test ». Ceci permet d'estimer à quel point le modèle sera généralisable.

L'underfitting, ou « sous-ajustement » du modèle, désigne une situation où l'apprentissage est insuffisant. Le modèle ne permet pas de prédire Y à partir de X de manière robuste et ses performances sont ainsi insuffisantes sur les données d'entrainement et sur les données de validation.

L'overfitting, ou « sur-ajustement » du modèle, désigne une situation où l'apprentissage est trop poussé, où le modèle est basé sur des caractéristiques trop spécifiques aux données d'entrainement, le rendant non généralisable, inexploitable pour de nouvelles données. Dans cette situation, les performances du modèle peuvent être très élevées sur les données d'entrainement mais vont s'effondrer une fois le modèle testé sur les données de validation.

Le risque d'overfitting est d'autant plus important que le modèle est complexe, que les paramètres du modèle sont nombreux, que les données d'entrainement sont bruitées et que le nombre d'échantillons est petit (figure 7). Ce risque dépend aussi de la méthode de machine learning utilisée. Les méthodes de sélection de variables permettent notamment de limiter le risque d'overfitting car elles permettent d'éliminer les variables redondantes ou non-discriminante et donc simplifier le modèle.

L'évaluation des performances des modèles sur les données de validation permet d'adapter certains paramètres d'apprentissage (ou « hyperparamètres ») pour limiter au maximum l'underfitting et l'overfitting. Cependant, pour vérifier que cette étape d'ajustement des

paramètres ne diminue pas la capacité de généralisation du modèle, il faudra l'évaluer sur des données supplémentaires, des données « test ».

Paradoxalement, dans le cas du deep learning le nombre très important de paramètres du modèle n'induit pas forcément d'overfitting. Ceci est lié au phénomène de « double descente » décrit avec le développement des réseaux de neurones artificiels [20] (figure 7). Il existe en réalité un seuil de complexité des modèles, à partir duquel le phénomène d'overfitting s'estompe.

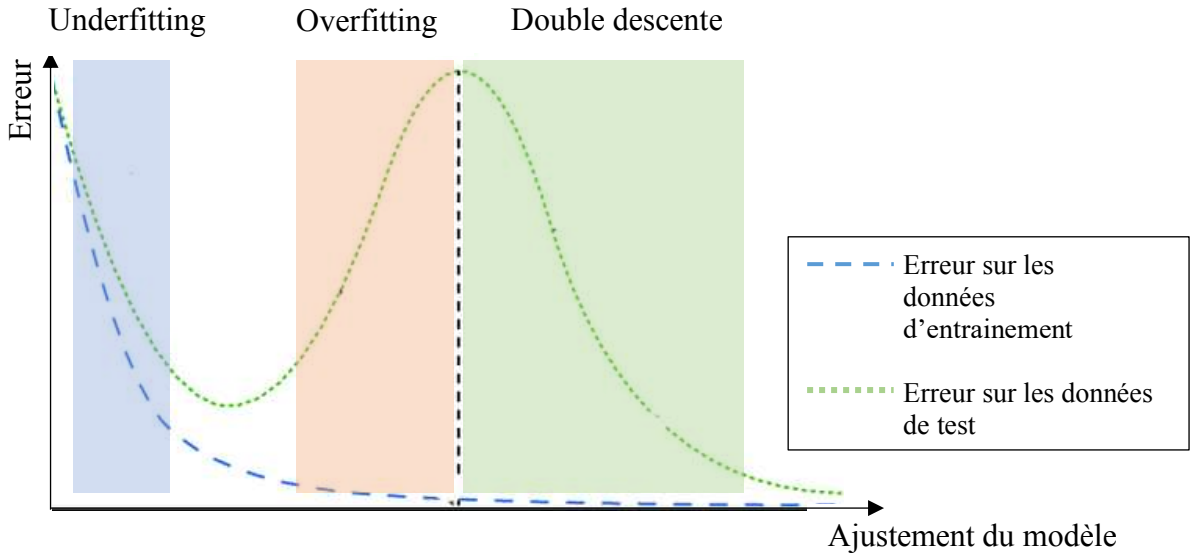

Figure 7. Représentation des phénomènes d'underfitting, d'overfitting et de double descente en apprentissage automatique.

E) Validation croisée (cross validation)

Plus la quantité de données utilisée pour l'entrainement d'un modèle est importante et plus l'entrainement sera efficace. De même, il est préférable de valider le modèle sur un nombre de données suffisant.

Cependant, il est fréquent de travailler avec un nombre limité d'observations dans les domaines de la biologie et de la santé. Ainsi, la technique de la validation croisée ou « cross validation » a été développée pour pouvoir utiliser les mêmes données pour l'entrainement et pour la validation du modèle.

Le principe est de répéter k fois l'étape de dichotomisation du jeu de données en jeu d'entrainement et jeu de validation, ainsi que l'entrainement du modèle et la mesure de ses performances. Chaque répétition est nommée « fold ». La base de données est divisée en k sous-

groupes. A chaque fold, un sous-groupe servira de base de validation et les autres serviront à l'entrainement. Au fold suivant, le sous-groupe suivant est utilisé pour la validation et le reste de la base est utilisée pour l'entrainement. Ainsi, une fois les k folds réalisés, chaque échantillon a été utilisé une fois pour la validation et k-1 fois pour l'entrainement (figure 8). Les performances du modèle peuvent ensuite être données en moyennant les résultats des k folds.

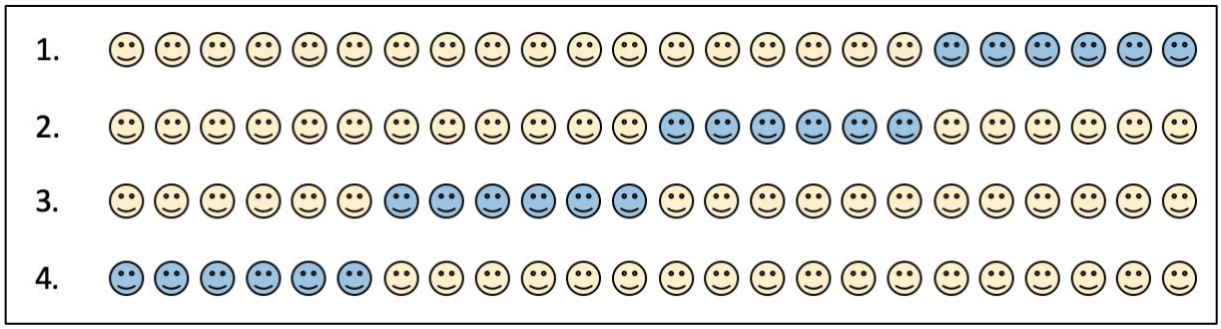

Figure 8. Exemple de validation croisée à 4 folds.

Pour chaque fold, les sujets jaunes sont utilisés pour l'entrainement et les sujets bleus sont utilisés pour la validation. Une fois les 4 folds terminés, chaque sujet a été utilisé 3 fois pour l'entrainement et une fois pour la validation

III. UTILISATIONS DU MACHINE LEARNING POUR L'ANALYSE DE DONNEES METABOLOMIQUES

Comme mentionné plus haut, les techniques de spectrométrie de masse utilisées en métabolomique non ciblée génèrent un grand nombre de spectres (correspondant à plusieurs dizaines de milliers d'ions) pour chaque échantillon analysé. Des traitements informatiques complexes de ces données de masse permettent d'individualiser quelques milliers de métabolites avec un certain nombre d'erreurs. Les méthodes de machine learning permettent d'analyser ces très nombreuses données dans leur ensemble.

Certaines méthodes de machine learning, en particulier de deep learning, ont été proposées pour automatiser le processus d'analyse informatique des données de masse, pour analyser les spectres de masse et ainsi individualiser les métabolites correspondants [21,22]. Cette application n'a pas été développée au cours de ce travail de thèse.

Les méthodes de machine learning sont également utilisées pour analyser les métabolites détectés après l'analyse métabolomique, et les relier au phénomène biologique étudié. Le plus souvent, ce phénomène biologique est défini à l'avance dans le schéma expérimental. Ainsi les méthodes les plus adaptées pour l'analyse statistique sont les méthodes d'apprentissage supervisé.

En revanche, **les méthodes d'apprentissage non supervisé** peuvent être utilisées pour identifier des échantillons correspondant à des sous-populations d'échantillons, présentant des caractéristiques métabolomiques similaires. Ainsi, dans un travail antérieur, nous avons pu comparer différentes méthodes d'apprentissage non supervisé pour identifier des sous-classes de tumeurs du sein (annexe 1) [16].

Les méthodes non supervisées permettent également de mettre en évidence des biais expérimentaux comme un effet batch ou « variabilité lot à lot », lié aux conditions de préparation des échantillons ou de paramétrage de l'analyse de LC-MS, ou d'identifier des « outliers », échantillons aux caractéristiques extrêmes ou aberrantes dans la population étudiée. Ainsi, en utilisant une analyse en composante principal (ACP), nous avons pu mettre en évidence un effet batch lors d'une étude portant sur les liens entre des données de métabolomique de tumeurs du sein et des données d'imagerie du métabolisme glucidique, par

tomographie par émission de positon au 18F-fluoro-2-deoxy-glucose (annexe 2) [23]. Dans le cas de cette étude, cet « effet batch », pourrait être due à un changement de colonne de chromatographie entre les analyses de deux lots d'échantillons.

Comme mentionné précédemment, les **méthodes d'apprentissage supervisé** restent les méthodes les plus fréquemment utilisées en métabolomique. Parmi ces méthodes, les méthodes les plus populaires sont, très largement, les méthodes PLS (Partial Least Square), puis les Random Forests (RF) et les SVM (Support Vector Machines) (Figure 9) [24]. Bien que très largement utilisés dans d'autre domaines, les réseaux de neurones artificiels (ANN : Artificial Neural Networks) sont encore peu courants en métabolomique. Les algorithmes génétiques (GA) ne sont presque plus utilisés.

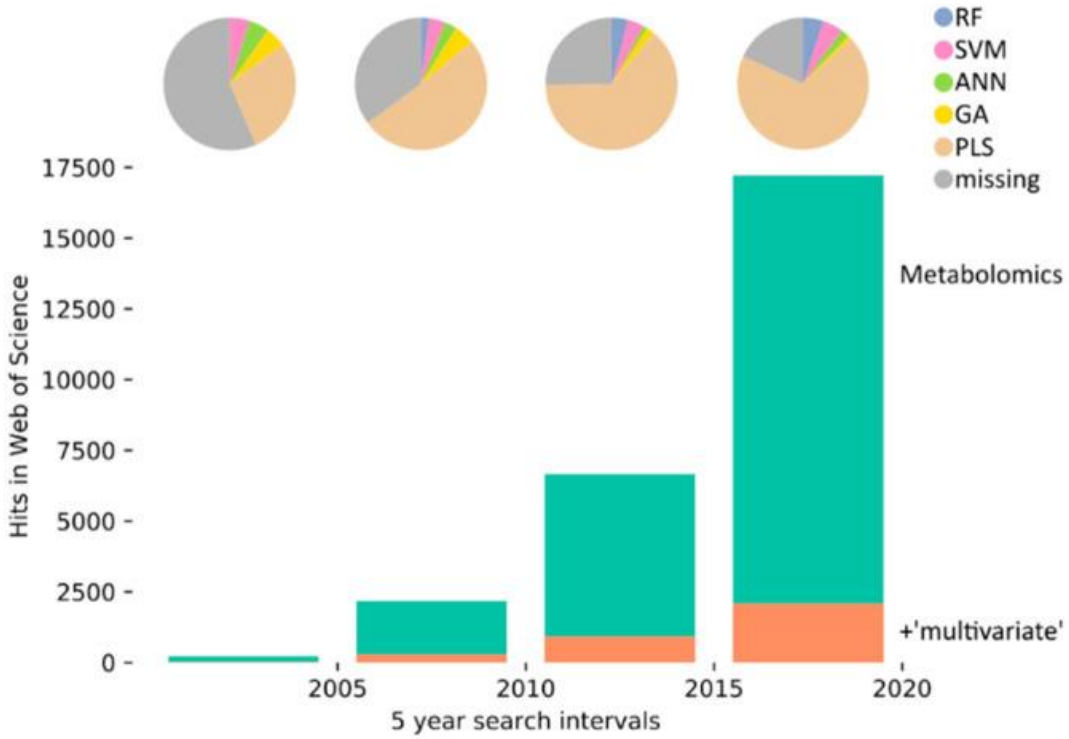

Figure 9. Nombres d'articles mentionnant 'metabolomics' (en vert) ou 'metabolomics' et 'multivariate' (orange) dans "Web of Science" sur des intervalles de 5 ans entre 2000 et 2020 (figure issue de l'article de Liebal et al. [24]). *Les diagrammes circulaires indiquent les proportions des méthodes utilisées dans les études de métabolomique. Les études de métabolomique sont en franche augmentation depuis les années 2000. Parmi les méthodes statistiques utilisées, les méthodes PLS : partial least squares (brun) sont franchement majoritaires. RF : Random Forrests, SVM : Standard Vector Machines, GA : Genetic Algorithms, ANN : Artificial Neural Networks, missing : non précisé ou non trouvé.*

A) Partial Least Squares – Discriminant Analysis (PLS-DA) et liens avec l'analyse en composante principale (ACP)

La PLS-DA [25] est une méthode de classification supervisée adaptée de la régression PLS, elle-même proche cousine de l'ACP [26]. Ces méthodes fonctionnent en réduisant le nombre de dimensions des variables explicatives via une analyse factorielle, créant des variables « latentes » ou « cachées » qui correspondent à des combinaisons linéaires de variables explicatives. Ces variables « cachées » sont appelées « composantes », ce qui explique la dénomination d'« analyse en composante principal » (ACP ou PCA en anglais) ou le fait que l'acronyme PLS soit parfois développé, en « Projection on Latent Space ».

Les composantes correspondent à des vecteurs dans un espace latent. Ces vecteurs u sont formés en multipliant chaque « variable explicative » X par une matrice de coefficients θ.

Dans le cadre de la PCA, pour chaque vecteur u les coefficients de la matrice θ sont choisis pour maximiser la variance des projections des variables observées X quand elles sont projetées sur le vecteur u (intuition : figure 10). La première composante sera créée en préservant au maximum la variance observée dans l'ensemble des données X. Ensuite, chaque nouvelle composante sera créée dans un plan orthogonal aux composantes précédentes, en préservant au maximum la variance résiduelle. Le nombre de composantes pouvant ainsi être créé n'est restreint que par le nombre de variables d'origine, cependant seules les 2 premières composantes sont utilisées en règle générale.

Ainsi, la PCA ne s'applique qu'aux variables explicatives X et ne prend pas en compte une éventuelle variable cible Y. On peut donc qualifier cette méthode d'apprentissage « non supervisé ». La PLS peut être décrite comme une variante « supervisée » de la PCA, qui prend en compte une variable cible Y et la covariance des variables X avec celle-ci. La régression PLS s'applique pour des variables Y quantitatives. En associant une régression PLS avec une analyse discriminante, la PLS-DA peut être appliquée à des variables Y qualitatives.

Si la PCA a pour but de réduire la dimension des données tout en préservant au maximum la variance des différentes variables X sous des contraintes de norme et d'orthogonalité des composantes, la PLS fait la même chose, mais en cherchant à préserver la covariance des

variables X avec Y. Ainsi, la régression PLS réalise un compromis entre la régression multiple de Y sur X et l'analyse en composantes principales de X.

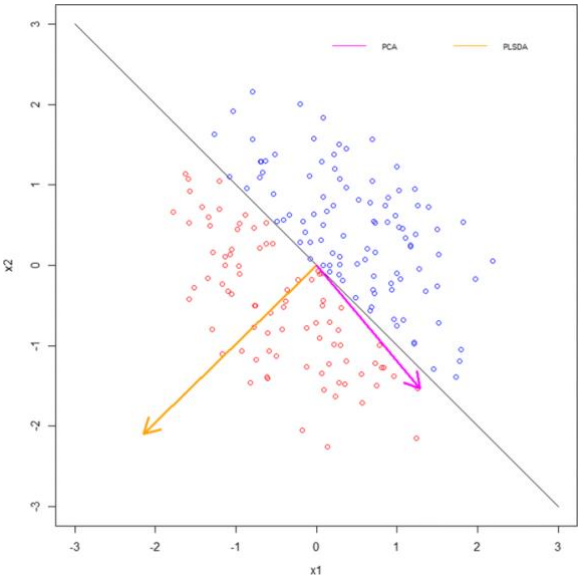

Figure 10 : Représentation des axes des premières composantes d'une PCA et d'une PLS-DA, appliquées à un jeu de données comprenant deux variables X (x1 et x2) et une variable Y (rouge vs bleu).
La composante principale de la PLS-DA (orange) suit un axe permettant de différencier les échantillons rouges et échantillons bleus alors que celle de la PCA (fuchsia) suit un axe qui ne le permet pas mais représente mieux la variance de x1 et x2.

La PCA est une méthode très utilisée en biologie[27,28], notamment parce qu'elle peut être appliquée quand le nombre de variables explicatives et largement supérieur au nombre d'échantillons (situation très fréquente en biologie) et parce qu'elle est associée à une représentation spatiale de la distribution des échantillons dans l'espace latent. La popularité de la PLS-DA en métabolomique est probablement liée, au moins en partie, à ses similarités avec la PCA.

Appliquer une régression (PLS-regression) ou une analyse discriminante (PLS-DA) aux composantes de la PLS permet de créer des modèles prédictifs, basés sur ces composantes.
Cependant aucune de ces méthodes n'inclue une étape de sélection de variable lors de la création du modèle prédictif. En effet, la PLS peut être utilisée pour réduire la dimension des données X, en résumant l'ensemble de X en quelques composantes principales. Mais le modèle prédictif final sera basé sur les composantes PLS, elles même issues de l'ensemble des données X. Ainsi, chaque variable de X aura un impact, parfois infime, sur le modèle final. De ce fait, la PLS-DA est sujette à l'overfitting [29].

De nombreux auteurs ont cependant détourné la PLS pour inclure une étape de sélection de variable [30,31], et ainsi limiter ce phénomène d'overfitting. Celle-ci se fait en général en sélectionnant les variables en fonction de leur importance pour la prédiction. La VIP (Variable

Importance in the Projection) permet de mesurer cette importance. Dans un modèle issu d'une PLS-DA incluant F composantes, basés sur J variables, la VIP de chaque variable *j* sera égale à la somme des poids W de *j* dans chaque composante $f(\sum_u W_{ju}^2)$, pondérée par le pourcentage de variance de Y expliqué par chaque composante $f(\mathrm{SSY}_f J/(\mathrm{SSY}_{tot.expl.}\mathrm{F}))$ (Formule 1)

$$\mathrm{VIP}_j^2 = \sum_f W_{jf}^2 \mathrm{SSY}_f J/(\mathrm{SSY}_{tot.expl.}\mathrm{F}) \qquad (1)$$

Ainsi la somme des VIP^2 de l'ensemble des J variables X sera égale à J, et la moyenne des VIP^2 sera égale à 1. On peut donc considérer que les variables ayant une VIP supérieure à 1 ont une importance supérieure à la moyenne pour la prédiction de Y. C'est pour cette raison que plusieurs auteurs utilisant la PLS-DA en métabolomique proposent de sélectionner les métabolites ayant une VIP supérieure à 1 [32,33].

Cependant, ce mode de sélection se révèle souvent insuffisamment discriminant car une VIP supérieure à 1 peut être purement liée à des aléas statistiques et ne traduit pas un résultat statistiquement significatif [34]. Ainsi, certains auteurs proposent un seuil de VIP plus élevé pour réaliser la sélection de variables[35]. En pratique, le choix du seuil de VIP pour cette étape de sélection de variable n'est pas bien défini et est souvent réalisé au cas par cas. Il en résulte un mode de sélection peu objectif, impliquant de ce fait un risque d'overfitting.

Des alternatives mieux standardisées existent pourtant pour réaliser cette étape de sélection de variable, notamment les méthodes incluant une régularisation L_1 (cf section III. F).

B) Arbres décisionnels et Random Forests

Le principe des arbres décisionnels est de dichotomiser les données en fonction des valeurs des variables explicatives pour prédire une variable cible. C'est une méthode intuitive ancienne qui a été utilisée pour créer de nombreuses méthodes plus complexes désignées communément comme les méthodes de classification et de régression par arbres décisionnels. Celles-ci ont notamment été développées par Breiman et al. en 1984 [36].

Un des avantages des arbres décisionnels réside dans leur facilité d'interprétation. De ce fait, ce type d'arbre est fréquemment utilisé hors du contexte du machine learning, notamment pour expliciter un mode opératoire pour la prise de décision en médecine (figure 11).

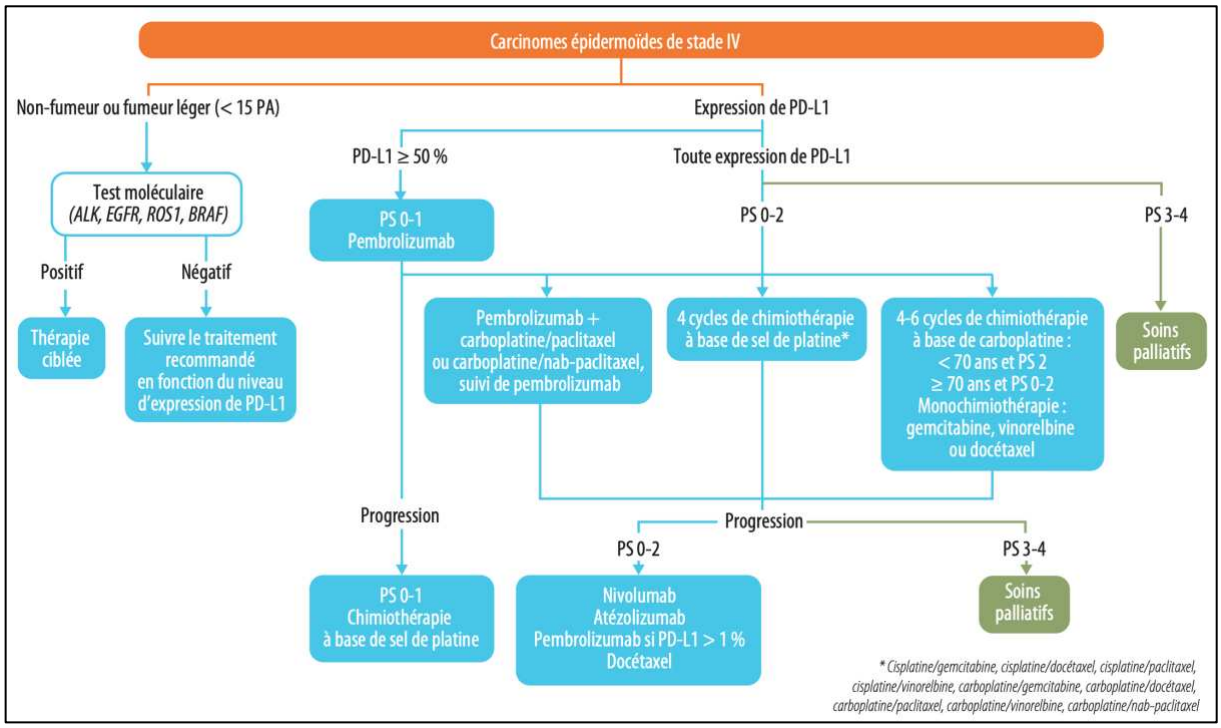

Figure 11. Exemple d'arbre décisionnel utilisé en médecine pour la prise en charge des carcinomes épidermoïdes pulmonaires de stade IV

Dans le cadre du machine learning, différents algorithmes existent pour entrainer les arbres décisionnels [36–39].Tous ces algorithmes se distinguent par le ou les critères de segmentation utilisés, par les méthodes d'élagages implémentées et par leur manière de gérer les données manquantes dans les prédicteurs. Si son apprentissage n'est pas limité, un arbre décisionnel aura toujours un ajustement parfait aux données d'entrainement et sera donc overfitté. Pour limiter l'overfitting il est possible d'utiliser des méthodes d'élagage. Le pré-élagage consiste à définir des critères d'arrêt de l'apprentissage à priori. Le post-élagage consiste à optimiser les performances de l'arbre en l'appliquant à une cohorte de validation après l'entrainement initial.

Breiman a proposé d'utilisé des forêts d'arbres décisionnels plutôt qu'un arbre décisionnel unique pour limiter l'overfitting et pallier au problème de sensibilité des arbres uniques à l'ordre des prédicteurs [40]. Le principe est d'entrainer un grand nombre d'arbres décisionnels sur plusieurs sous-ensembles des données d'entrainement, en utilisant la cross validation pour limiter le surentrainement des arbres. Une fois l'ensemble des arbres décisionnels entrainés, les

prédictions sont faites selon les prédictions de l'ensemble des arbres, par vote majoritaire (figure 12).

Comme mentionné précédemment, l'avantage de cette méthode est de limiter le risque d'overfitting. L'inconvénient est que le processus de prédiction n'est plus intuitivement interprétable.

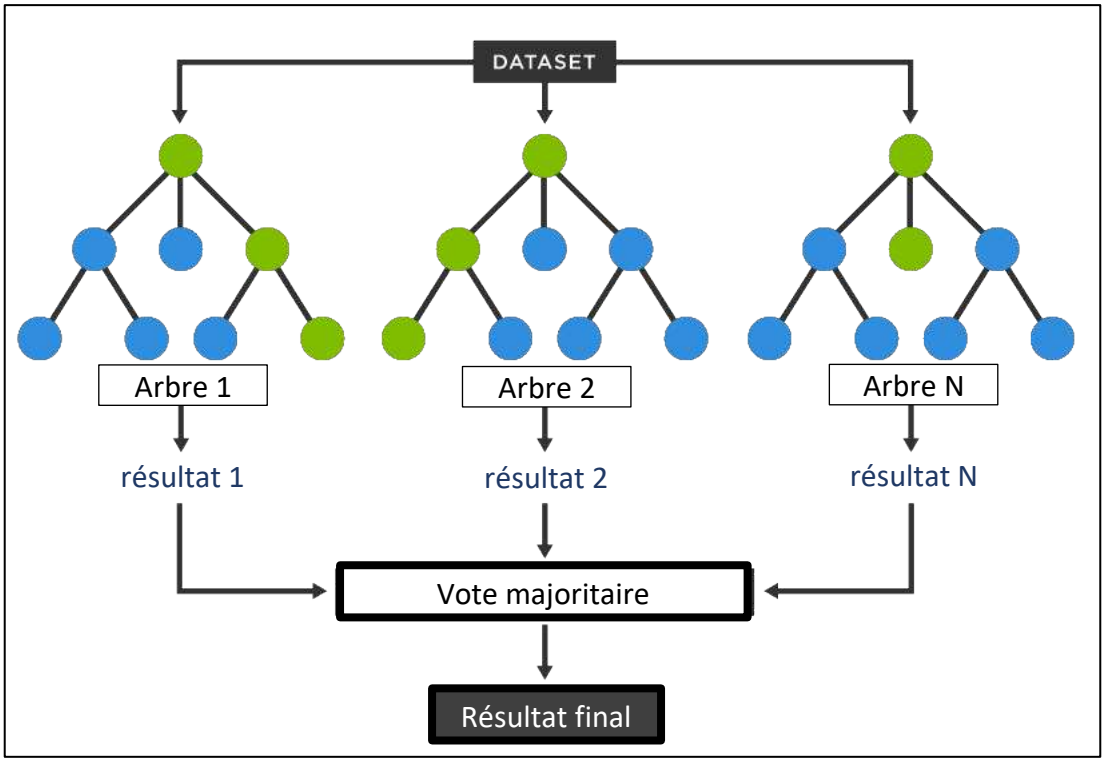

Figure 12. Représentation schématique de la méthode Random Forest

C) Machines à vecteurs de support ou « Support Vector Machines » (SVM)

Les machines à vecteurs de support ou « Support Vector Machines » (SVM) sont des méthodes d'apprentissage supervisé développées par Vapnik et al. [41,42]. En 1963, Vapnik et al. ont proposé des SVM linéaires. Le principe est de rechercher un hyperplan linéaire dans un espace de dimension égale au nombre de variables explicatives x pour séparer au mieux les classes y (Figure 13). On pourra prédire à quelle classe appartient un nouveau sujet en fonction du côté de l'hyperplan dans lequel il se trouve.

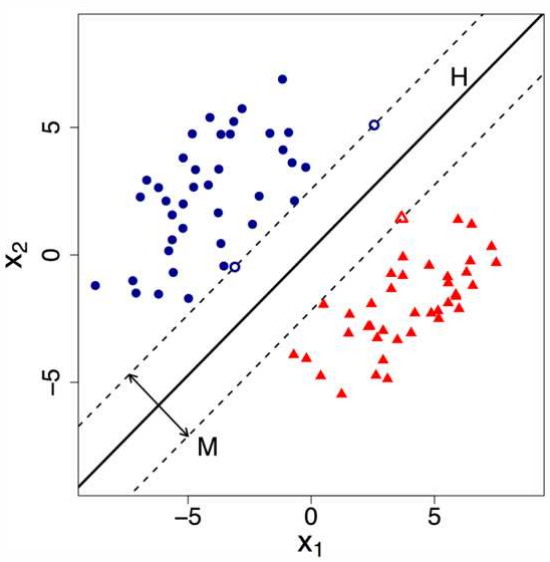

Figure 13. Représentation graphique d'un SVM pour une classification binaire dans un espace bidimensionnel.

H correspond à l'hyperplan de séparation optimal. Les ronds bleus et les triangles rouges représentent les deux classes. Les sujets les plus proches H sont nommés « vecteurs de support ». M correspond aux marges entre H et les vecteurs de supports.

La méthode initialement proposée en 1963 ne fonctionne pas pour des données non séparables linéairement car elle utilise des marges dites « rigides » et ne vise qu'à maximiser la distance entre ces marges. Dans ces conditions, s'il n'existe pas d'hyperplan H séparant complètement les données, l'algorithme ne peut aboutir. Dans les années 1990, Vapnik et al. ont proposé une méthode SVM à marges « souples », permettant une utilisation même pour des données non séparables linéairement. Pour cela l'algorithme ne cherche plus uniquement à maximiser la distance entre les marges M mais à trouver un compromis entre la maximisation des marges et la minimisation du nombre de sujets mal classés. Ensuite, Bosner, Guyon et Vapnik ont proposé une méthode SVM non linéaires, en utilisant l'astuce du noyau (kernel trick en anglais). Le principe est de projeter les données d'un espace dans lequel elles ne sont pas linéairement séparables dans un espace dans lequel elles le sont. Les dimensions supplémentaires sont obtenues en appliquant une fonction noyau aux données.

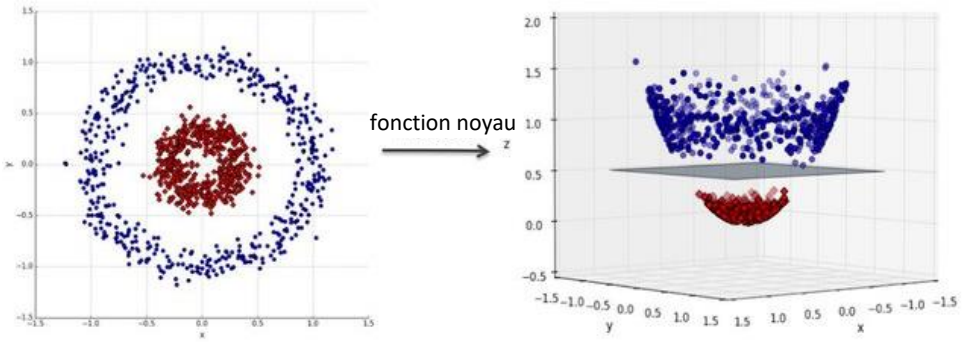

Figure 14. Représentation de l'astuce du noyau. *En 2D, les classes rouges et bleus, réparties en cercles, ne sont pas linéairement séparables. Si on ajoute une troisième dimension représentative de la « distance au centre », tel que (x,y) devient (x, y, $x^2 + y^2$), les classes deviennent linéairement séparables dans l'espace 3D.*

D) Algorithmes génétiques ou « Genetic Algorithms » (GA)

Les algorithmes génétiques ont été initialement développés dans les années 1970 par John Holland [43]. Leur principe est d'utiliser les principes de l'évolution décrits par Darwin et en génétique pour trouver des solutions à des problèmes de classification ou de régression. L'idée générale est de démarrer l'entrainement avec un certain nombre de solutions potentielles nommées « individus ». Par analogie avec le domaine de la génétique, les solutions ou « individus » sont composées de « chromosomes », contenant eux-mêmes des gênes.

Un premier set d'individus est généré aléatoirement. Les individus sont évalués via une fonction d'erreur puis les plus adaptés au problème sont sélectionnés et sont utilisés pour créer de nouveaux individus par fusion des chromosomes, avec application de phénomènes de croisement et de mutation des gènes dans ces chromosomes. Ainsi, une nouvelle population est générée, légèrement mieux adaptée au problème que la population initiale. Le processus est répété un grand nombre de fois jusqu'à un critère d'arrêt, basé sur la fonction d'erreur (Figure 15).

Ces méthodes ont l'avantage de permettre de trouver une solution proche de l'optimum pour des problèmes variés, linéaires ou non. Cependant, elles nécessitent un grand nombre de calculs, leur succès dépend grandement de la population initialement générée et des paramètres de croisement et de mutation, et elles ne permettent pas d'être certain d'avoir atteint la meilleure solution au problème étudié.

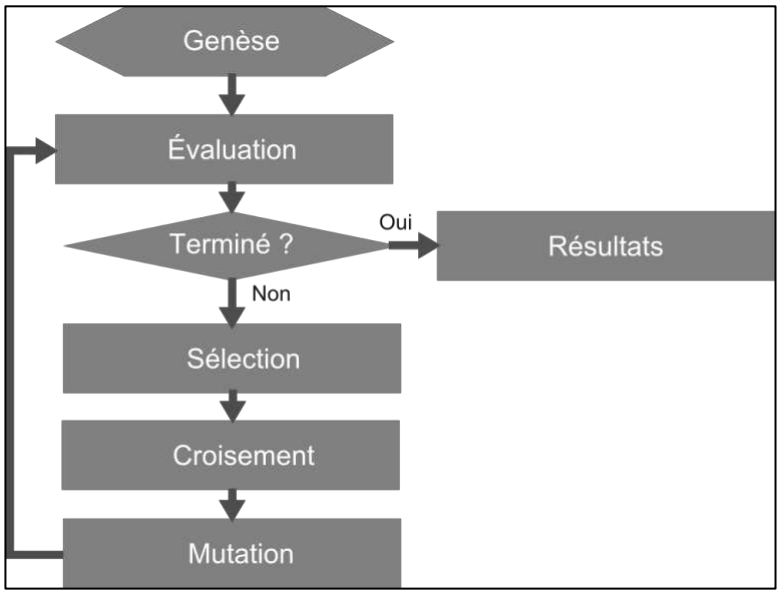

Figure 15. Représentation schématique d'un algorithme génétique

E) Réseaux de Neurones Artificiels ou « Artificial Neural Networks » (ANN)

Les réseaux de neurones artificiels ou « artificial neural networks » (ANN) correspondent à un type de modèle mathématique et informatique, inspiré des circuits neuronaux biologiques. Les ANN sont constitué d'un réseau de neurones formels, tels que décrit par McCulloch et Pitts en 1943 [44] (figure 16).

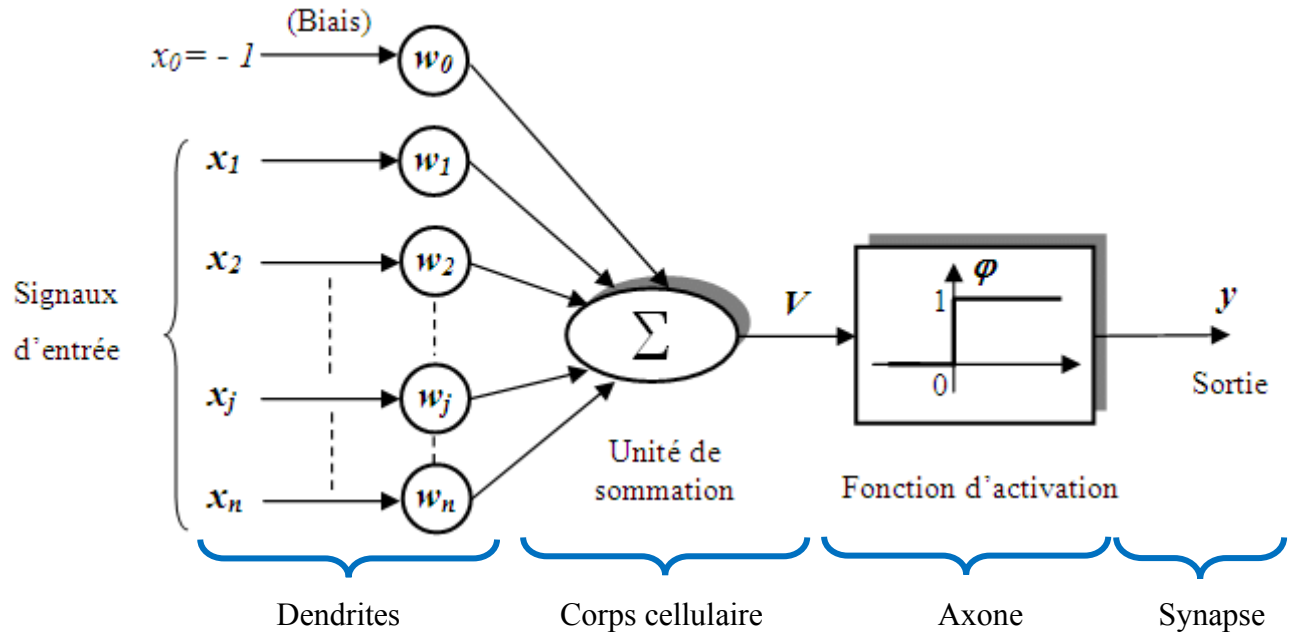

Figure 16. Représentation du neurone formel, tel que décrit par McCulloch et Pitts.

Un neurone unique peut être utilisé pour réaliser une tache de classification binaire. C'est ce que Rosenblatt a nommé « perceptron » [45]. L'association de plusieurs neurones formels permet de réaliser des taches de classification multi-classes, toujours avec des limites linéaires. L'organisation de nombreux neurones formels en réseau multi-couches (figure 17) permet de réaliser des taches de classification multi-classes avec des limites non linéaires. Un réseau de neurones multi-couches est composé d'une succession de couches de neurones, dont chacune prend ses entrées sur les sorties de la précédente. La notion d'apprentissage « profond » fait référence à l'utilisation de couches « cachées » dans ces ANN multi-couches.

Les poids w de chaque neurones sont modifiés lors de la phase d'apprentissage grâce à un algorithme de rétro-propagation du gradient [46].

De nombreux types de réseaux de neurones existent. Ceux-ci varient en fonction de la topologie des connexions neuronales (architecture du réseau), des fonctions utilisées comme unités de sommation, des fonctions d'activation et des types de sortie dans chaque couche.

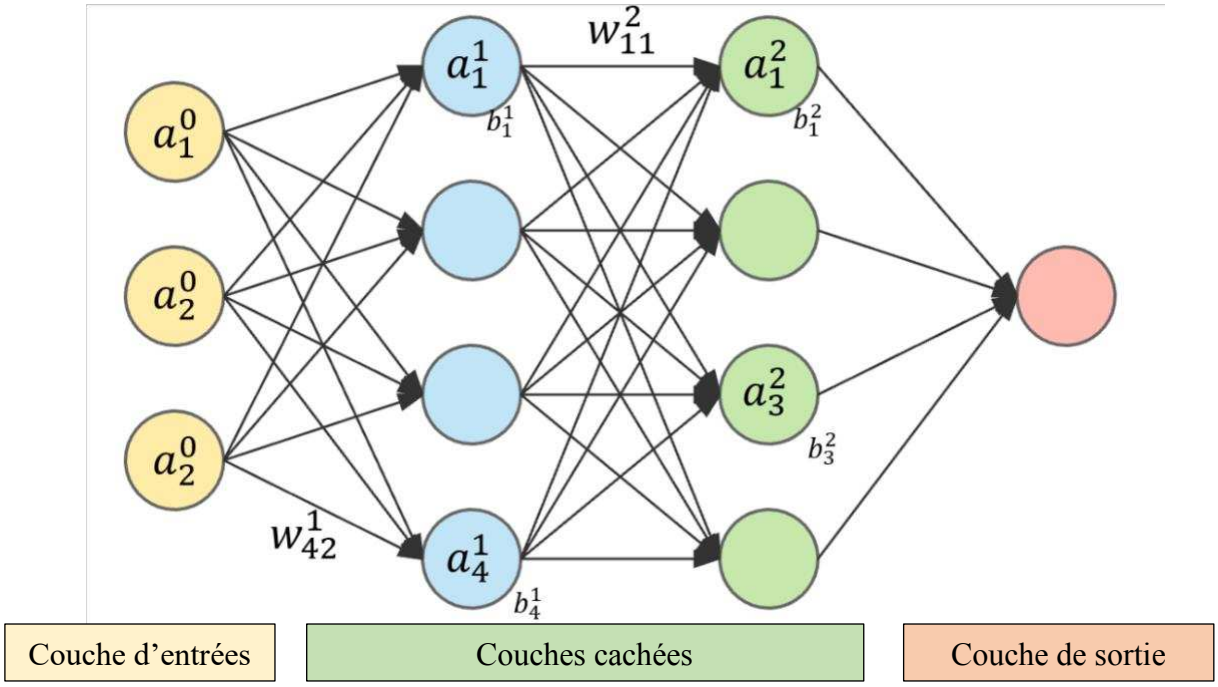

Figure 17. Représentation schématique d'un réseau de neurones multi-couches.

F) Méthodes de régularisation

La régularisation fait référence à un processus consistant à ajouter une pénalité à un problème, notamment pour éviter le sur-apprentissage. La méthode la plus généralement utilisée est de pénaliser les valeurs extrêmes des paramètres, qui correspondent souvent à un sur-apprentissage. Pour cela, une norme est ajoutée à la fonction d'erreur à minimiser. Les normes les plus couramment employées sont les normes L_1 et L_2.

- La norme L_1 prend la forme : $erreur\ après\ régularisation = erreur + \alpha \sum |W|$
- La norme L_2 prend la forme : $erreur\ après\ régularisation = erreur + \alpha \sqrt{\sum W^2}$

L'ajout d'une norme aura pour effet d'empêcher les poids de prendre des valeurs extrêmes (ce qui est souvent lier à un overfitting). Quand une norme L_1 est appliquée, certains poids seront

mis à zéro tandis que quand une norme L_2 est appliquée certains poids tendront vers des petites valeurs, sans atteindre zéro (figure 18). Ainsi L_1 offre l'avantage d'induire une sélection de variable.

Les méthodes de régularisation peuvent s'appliquer à de nombreuses méthodes de machine learning. La régularisation L1 est également nommée régression LASSO (pour Least Absolute Shrinkage and Selection Operator) et la régularisation L2 est également appelée régression « Ridge ».

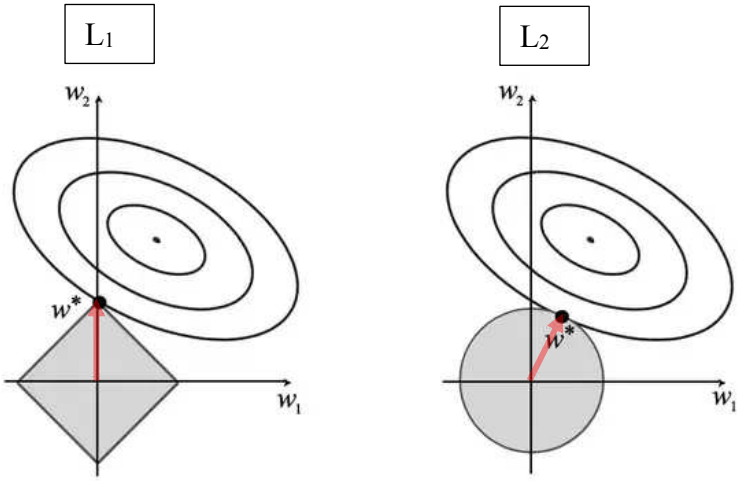

Figure 18. Représentation schématique de la différence entre la norme L_1 et la norme L_2. *La portion grisée représente la zone de contrainte, les ellipses représentent la fonction d'erreur. L'objectif est de se rapprocher au maximum de l'erreur minimale atteignable (ellipse au contact de la zone de contrainte). En appliquant une norme L_1, le poids W_2 va augmenter tandis que W_1 vas rester à 0, alors qu'en appliquant L_2, les deux poids vont augmenter.*

G) Comparaison inter-méthodes

Ainsi, chaque méthode présente des avantages et des inconvénients [47]. Le plus souvent en pratique, plusieurs méthodes sont testées et les résultats sont confrontés les uns aux autres pour élaborer des conclusions. De même, certains auteurs ont développé des outils permettant de tester de nombreuses méthodes en une seule analyse [48]. Il est fréquent que plusieurs méthodes présentent des performances similaires pour un même jeu de données [49]. Dans ce cas, le choix de la méthode repose en général sur sa simplicité. A performances égales, un modèle linéaire simple sera plus intéressant qu'un modèle non linéaire complexe car le premier sera plus facilement interprétable. Les résultats dépendront quoi qu'il en soit de la qualité des données analysées. Il sera plus probable de développer un modèle performant avec des données peu bruitées et concernant un grand nombre d'échantillons.

IV. DEVELOPPEMENT DE DEUX NOUVELLES METHODES DE CLASSIFICATION SUPERVISEE UTILISABLES POUR L'ANALYSE DE DONNEES DE METABOLOMIQUE

Dans le cadre de cette thèse, nous avons participé au développement et l'évaluation de deux nouvelles méthodes de machine learning en collaboration avec des mathématiciens du laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis (I3S) : une méthode utilisant une méthode Primal-Dual [50,51] (premier article) et une autre basée sur un auto-encoder supervisé (deuxième article). Ces méthodes d'analyse ont été adaptées pour répondre à des besoins majeurs de l'utilisation des données omiques dans un contexte médical, et notamment de la métabolomique : la fiabilité des résultats et leur interpérabilité.

Fiabilité : intérêt d'un indice de confiance

Tout d'abord, pour qu'un modèle prédictif soit suffisamment fiable pour être utilisable en médecine, il faut que ses performances diagnostiques (exactitude, sensibilité, et spécificité) soient élevées. Les nouvelles méthodes abordées ont donc été optimisé pour obtenir les meilleurs performances diagnostiques possibles.

Toutefois, selon nous, les mesures de performance basées sur des prédictions univoques, représentent mal la pratique courante. En pratique il existe en général une part d'incertitude dans la démarche diagnostique. Lorsqu'un praticien pose un diagnostic, il y associe intuitivement une estimation de la probabilité sous-jacente. Ainsi, en fonction des situations, certains diagnostics sont posés avec certitude quand d'autres sont plus incertains.

Dans le cadre des nouvelles méthodes de machine learning proposées, nous avons inclus un « score de probabilité » ou « indice de confiance » associé à chaque prédiction. Celui-ci nous semble particulièrement pertinent pour une utilisation en pratique clinique car il permet à l'utilisateur d'évaluer la probabilité du résultat obtenu et de confronter ce résultat aux autres données disponibles pour chaque patient.

Représentation graphique simplifiée du résultat

Par ailleurs, l'auto-encodeur supervisé inclut également une représentation graphique de la distribution des échantillons dans l'espace latent, similaire aux représentations graphiques associées à la PLS-DA ou à l'ACP. Cette représentation graphique est généralement appréciée

dans les analyses de classification car elle permet une représentation globale des résultats. Elle permet ainsi d'évaluer visuellement la bonne séparation des groupes étudiés, reflet des performances du modèle et donc de la fiabilité des résultats.

Interprétabilité du résultat

Pour utiliser un modèle prédictif en médecine il paraît pertinent de pouvoir comprendre son fonctionnement. Limiter le nombre de variables intégrées dans le modèle et permettre de quantifier leur impact respectif pour la prédiction finale facilite cette compréhension. Les nouvelles méthodes proposées dans cette thèse incluent une étape de sélection de variable, limitant le nombre de variables intégrées dans le modèle final. De plus, l'importance de chaque variable dans le modèle est directement donnée sous forme de poids, quantification la plus intuitive possible de l'impact de la variable pour la prédiction.

Dans le cadre de tests diagnostiques basés sur des données de métabolomique, pouvoir facilement identifier les métabolites pertinents et quantifier leur impact pour la prédiction permet d'étudier les relations biologiques pouvant exister entre ces métabolites et les pathologies étudiées. En pratique, comprendre le rationnel biologique sous-tendant le modèle prédictif permet de mieux appréhender les prédictions réalisées.

A) Principes

La méthode PD-CR : Primal Dual for Classification with regression

Le principe de dualité désigne le principe selon lequel les problèmes d'optimisation peuvent être vus de deux perspectives : un problème primal et un problème dual. La solution du problème dual donne une borne inférieure à la solution du problème de minimisation primal. En général les valeurs optimales des problèmes primal et dual ne sont pas forcément égales : la différence est appelée « saut de dualité ». Pour les problèmes en optimisation convexe, ce saut est nul sous contraintes.

Barlaud et al. [52] ont proposé une nouvelle méthode de classification, combinant une étape de réduction de dimensions par projection sur un espace latent et une étape de sélection de variable par régularisation l_1. Pour se faire, ils proposent de résoudre un problème de classification supervisée convexe et contraint qu'ils posent ainsi :

$$\min_{(W,\mu)} \|Y\mu - XW\|_1 + \frac{\rho}{2}\|I_k - \mu\|_F^2 \text{ s.t. } \quad \|W\|_1 \leqslant \eta$$

Où Y correspond à la matrice labels de forme $\{0, 1\}^{m\times k}$, X à la matrice de variables explicatives de dimension $m \times d$, W aux poids des variables, μ aux centroïdes de chaque classe dans l'espace latent, I_k la matrice identité de dimension k et η le rayon de la boule l_1. W et μ sont inconnus. Un terme de régularisation l_2 est ajouté, où ρ est un hyperparamètre, pour limiter la solution triviale μ=0 et W=0.

Ils proposent de solutionner ce problème en le dualisant sous la forme :

$$\min_{(W,\mu)} \max_{\|Z\|_\infty \leqslant 1} \langle Z, Y\mu - XW\rangle + \frac{\rho}{2}\|I_k - \mu\|_F^2 \text{ s.t. } \|W\|_1 \leqslant \eta$$

Où Z est une matrice de dimension $m \times k$.

Ils utilise pour cela un algorithme inspiré de celui proposé par Condat [53] :

Input: $X, Y, N, \sigma, \tau, \eta, \rho, \mu_0, W_0, Z_0$
$W := W_0$
$\mu := \mu_0$
$Z := Z_0$
for $n = 1, \ldots, N$ **do**
  $W_{\text{old}} := W$
  $\mu_{\text{old}} := \mu$
  $W := W + \tau \cdot (X^T Z)$
  $W := \text{proj}(W, \eta)$
  $\mu := \frac{1}{1+\tau_\mu \cdot \rho}(\mu_{\text{old}} + \rho \cdot \tau_\mu I_k - \tau_\mu \cdot (Y^T Z))$
  $Z := \max(-1, \min(1, Z + \sigma \cdot (Y(2\mu - \mu_{\text{old}}) - X(2W - W_{\text{old}}))))$
end for
Output: W, μ

Auto-encodeur supervisé non paramétrique

Les auto-encodeurs sont des réseaux de neurones artificiels dont l'architecture a initialement été pensée pour réduire la dimension de bases de données et/ou les débruiter. Pour se faire les données X sont encodées via un certain nombre de couches cachées, jusqu'à prendre une forme Z, matrice de moindre dimension composée de données latentes, qui est ensuite décodée via un nombre de couches cachées égal au nombre de couches d'encodage, pour donner une matrice X' de dimension égale à la matrice X d'entrée. Les paramètres des différentes couches cachées sont appris pour minimiser la différence entre X et X' (figure 19). A noter que si l'encodeur ne comporte qu'une couche cachée appliquant des transformations linéaires, celui-ci donnera un résultat très proche d'une ACP.

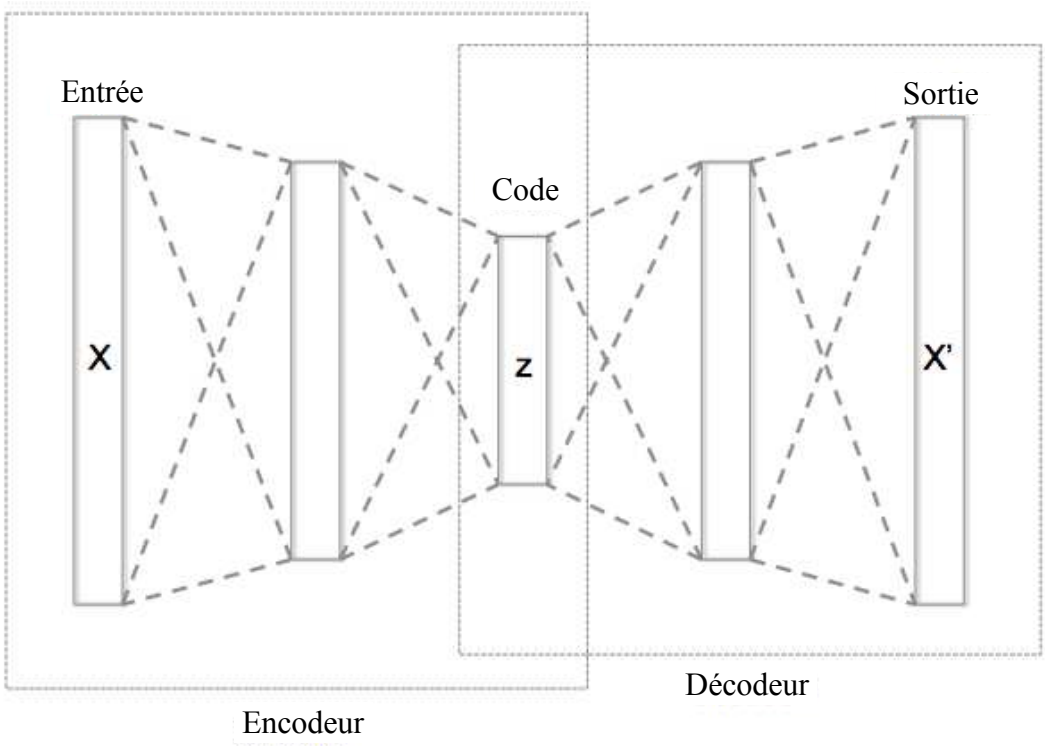

Figure 19. Représentation de l'architecture d'un auto-encodeur.

Certains auteurs ont proposé d'utiliser cette architecture pour générer de nouvelles données [54]. En effet, une fois l'encodeur et le décodeur entrainés, l'application du décodeur à un point aléatoire de l'espace latent Z devrait générer un point dans l'espace X', proche des observations initiales de l'espace X.

Cependant, en cas d'overfitting, l'organisation de l'espace latent, trop spécifique aux données d'entrainement X, rendra les données générées aberrantes et donc inexploitables.

Pour limiter l'overfitting lors de la phase d'apprentissage de l'auto-encodeur, Kingma et Welling ont proposé de contraindre la distribution des variables dans l'espace latent [55] . Ces autoencodeurs induisant une distribution prédéfinie dans l'espace latent sont nommés Variational AutoEncoders (VAE). En règle générale, une distribution gaussienne est choisie, notamment pour faciliter le processus informatique d'entrainement du VAE. Cependant, forcer une représentation gaussienne des données dans l'espace latent peut induire un biais si la structure réelle des données n'est pas gaussienne.

Barlaud et al. [56] ont proposer un auto-encodeur pour lequel la distribution dans l'espace latent n'est pas influencée par une distribution gaussienne à priori mais par la classification des données. Ainsi la distribution initiale des données est conservée et la distribution dans l'espace latent reste structurée, permettant la génération de données. Barlaud et al. proposent d'entrainer leur auto-encoder en minimisant une fonction d'erreur basée à la fois sur l'erreur de l'étape de reconstruction et sur l'erreur de classification, tout en incluant un terme de régularisation (figure 20).

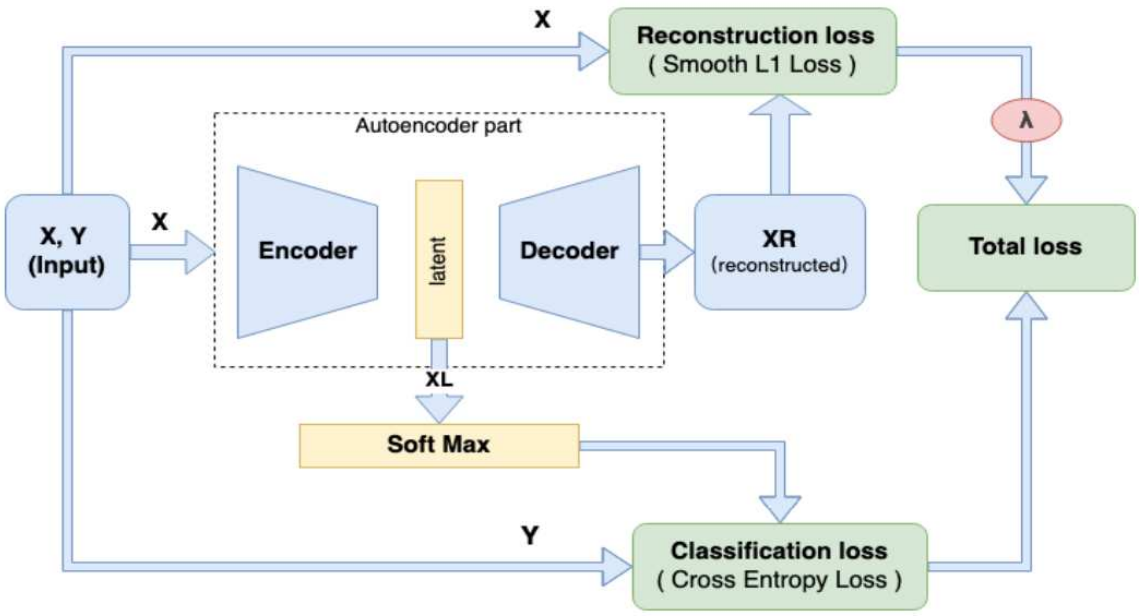

Figure 20. Architecture de l'autoencodeur supervisé non paramétrique proposé par Barlaud et al. [56]

Ainsi, les poids W de l'encodeur et du décodeur sont appris pour que XR soit le plus proche possible de X, tout en faisant en sorte que la classification de XL (X projeté dans l'espace latent) soit le plus proche possible de Y.

Cette méthode combine ainsi une étape de réduction de dimension par projection des données X dans un espace latent de moindre dimension, une étape de sélection de variable par l'adjonction d'un terme de régularisation et une étape de classification supervisée.

B) Applications pratiques

Dans le cadre de cette thèse, nous avons participé à l'implémentation de ces méthodes en langage python pour une utilisation sur des données omiques.

Nous avons testé ces méthodes sur des données de métabolomique et nous les avons comparées aux méthodes les plus fréquemment utilisées : PLS-DA, Random Forests et SVM. Ainsi, nous avons pu montrer que ces méthodes permettaient d'obtenir des modèles au moins aussi fiables, voire plus fiables que les méthodes classiquement utilisées en métabolomique.

Ces méthodes ont été implémentées en langage python et sont librement disponibles sur internet (https://github.com/CyprienGille/Supervised-Autoencoder et https://github.com/tirolab/PD-CR). Cependant, certains utilisateurs peuvent préférer un format plus intuitif, tels que celui proposé pour les méthodes de Metaboanalyst.ca [57]. Nous développons actuellement une interface plus intuitive pour une meilleure distribution de ces méthodes.

Il est difficile de prédire à priori la meilleure méthode d'analyse pour une série de données de métabolomique. Celle-ci dépend de la question posée mais également de la structure des données. Dans ce contexte, ces deux nouvelles méthodes viennent étoffer le panel disponible pour les études de métabolomique.

Article 1 : Primal-dual for classification with rejection (PD-CR): a novel method for classification and feature selection-an application in metabolomics studies

David Chardin, Olivier Humbert, Caroline Bailleux, Fanny Burel-Vandenbos, Valerie Rigau, Thierry Pourcher, Michel Barlaud

BMC Bioinformatics. 2021 Dec 15;22(1):594. doi: 10.1186/s12859-021-04478-w.

Primal-Dual for Classification with Rejection (PD-CR): A novel method for classification and feature selection. An application in metabolomics studies.

David Chardin, Olivier Humbert, Caroline Bailleux, Fanny Burel-Vandenbos, Valerie Rigau, Thierry Pourcher, Michel Barlaud

15/12/2021

Abstract

Background: Supervised classification methods have been used for many years for feature selection in metabolomics and other omics studies. We developed a novel primal-dual based classification method (PD-CR) that can perform classification with rejection and feature selection on high dimensional datasets. PD-CR projects data onto a low dimension space and performs classification by minimizing an appropriate quadratic cost. It simultaneously optimizes the selected features and the prediction accuracy with a new tailored, constrained primal-dual method. The primal-dual framework is general enough to encompass various robust losses and to allow for convergence analysis. Here, we compare PD-CR to three commonly used methods : Partial Least Squares Discriminant Analysis (PLS-DA), Random Forests and Support Vector Machines (SVM). We analyzed two metabolomics datasets: one urinary metabolomics dataset concerning lung cancer patients and healthy controls; and a metabolomics dataset obtained from frozen glial tumor samples with mutated isocitrate dehydrogenase (IDH) or wild-type IDH.

Results: PD-CR was more accurate than PLS-DA, Random Forests and SVM for classification using the 2 metabolomics datasets. It also selected biologically relevant metabolites. PD-CR has the advantage of providing a confidence score for each prediction, which can be used to perform classification with rejection. This substantially reduces the False Discovery Rate.

Conclusion: PD-CR is an accurate method for classification of metabolomics datasets which can outperform PLS-DA, Random Forests and SVM while selecting biologically relevant features. Furthermore the confidence score provided with PD-CR can be used to perform classification with rejection and reduce the false discovery rate.

1 Introduction

Among the different omics fields, metabolomics is the most recent and provides new insights for a global study of biological systems. Metabolomics is a rapidly growing and promising field of research in biology and healthcare. Metabolomics approaches are based on the determination of the levels of different small molecules or metabolites in biological samples (tissue, cells, serum, urine...). Interestingly, ever since the early metabolomics studies, supervised classification methods have been used for the analysis of the related datasets. One of the initial aims of metabolomic studies was to establish useful biomarkers, indicative of specific physiological states or aberrations. The challenge now is to understand the mechanisms by which changes in the metabolome are implicated in different phenotypic outcomes in a complex systems biology approach [1, 2].

Most metabolomics studies generate complex multivariate datasets including varying correlations between features and systematic noise. Therefore, multivariate data analysis methods are needed to explore these datasets. One of the most frequently used methods for metabolomics analyses is Partial Least Squares-Discriminant Analysis (PLS-DA) [3, 4].

PLS-DA is a chemometric technique used to optimize separation between different classes of samples, which is accomplished by linking two data matrices: X (raw metabolomic data) and Y (class membership). It has the advantage of handling highly collinear and noisy data. Yet, it has some drawbacks and needs to be handled with caution. Indeed it has been reported that PLS-DA can: 1. Lead to over-fitting when the number of variables significantly exceeds the number of samples. Indeed, in this setting, the model is likely to lead to accurate classification by chance, based on irrelevant features [5]; 2. Have difficulties when few variables are responsible for the separation between two or more classes and, therefore, require a larger number of variables to achieve a good prediction accuracy [6]; and finally, 3. Lead to an over-optimistic understanding of the separation between two or more classes [7].

Continuous effort is being made to provide new statistical tools to tackle these drawbacks [8]. Some authors use Random Forests [9] as an alternative to PLS-DA for metabolomics studies [10]. Random Forests are based on the bagging algorithm and use an Ensemble Learning technique. Random Forests create a large number of decision trees and combine their outputs. Yet, Random Forests have significant drawbacks. For instance, they tend to over-fit when using noisy datasets. Furthermore, the main disadvantage of Random Forests is their complexity. Indeed, they are much harder and time-consuming to construct, require more computational resources and are less intuitive than decision trees. Furthermore this complexity significantly hampers their interpretability. Support Vector Machines (SVM) are another option [11, 12] but have similar drawbacks as

Random Forests and are particularly consuming in computational resources.

Mathematics I3S partner has recently introduced a new tailored, constrained primal-dual method for supervised classification and feature selection [13]. This method has the significant advantage of providing a trustworthy confidence index with each prediction, which we use to define a new classifier with rejection. This is particularly useful in the context of clinical decision making as it diminishes the number of false positive and false negative results. Moreover, we believe this method out-performs other methods in terms of accuracy and feature selection.

Although there are many machine learning methods for feature selection such as LASSO [14, 15], Discriminant analysis [16], Proximal methods [17, 18] and Boosting [19, 20], here we compare our novel Primal-Dual method for Classification with Rejection (PD-CR) to the state of the art PLS-DA and Random Forests and SVM classification methods frequently used in metabolomics studies.

2 Methods

2.1 Mathematical background

2.1.1 Robust classification and regression using ℓ_1 centers

Mathematically, classification problems can be described as follows :
Let X be the $m \times d$ data matrix made of m line samples $x_1, \ldots, x_m$ that belong to the d-dimensional space of features.
Let $Y \in \{0,1\}^{m \times k}$ be the matrix of labels where $k \geq 2$ is the number of clusters. Each line of Y has exactly one nonzero element equal to one, $y_{ij} = 1$ indicating that the sample x_i belongs to the j-th cluster. Projecting the data in lower dimension is crucial to be able to separate them accurately.
Let W be the $d \times k$ projection matrix, where $k \ll d$. (Note that the dimension of the projection space is equal to the number of clusters.)
The goal of the supervised classification method is to find the best possible values for the projection matrix W.

Sparse learning based methods have received a lot of attention in the last decade because of their high level of performance. The basic idea is to use a sparse regularizer that forces some coefficients to be zero. To achieve feature selection, the *Least Absolute Shrinkage and Selection Operator* (LASSO) formulation [14, 21, 22, 23, 24, 25] adds an ℓ_1 penalty term to the classification cost. An accurate criterion is based on

the sum of the square difference (used in k-means [26]) and can be cast as follows:

$$\|Y\mu - XW\|_F^2 = \sum_{j=1}^{k} \sum_{l \in C_j} \|(XW)(l,:) - \mu_j\|_2^2, \tag{1}$$

where $C_j \subset \{1, \ldots, m\}$ denotes the j-th class, and where the row vector μ_j is the centroid of this class. Therefore, the matrix of centers μ is a square matrix of order k. It is well known that the Frobenius norm is sensitive to outliers. To address this, we have improved the approach by replacing the Frobenius norm by the ℓ_1 norm of the loss term as follows :

$$\|Y\mu - XW\|_1 = \sum_{j=1}^{k} \sum_{l \in C_j} \|(XW)(l,:) - \mu_j\|_1. \tag{2}$$

where $C_j \subset \{1, \ldots, m\}$ denotes the j-th cluster, and where $\mu_j := \mu(j,:)$ is the j-th line of μ. In our method, we simultaneously optimize (W, μ), adding some *ad hoc* penalty to break homogeneity and avoid the trivial solution $(W, \mu) = (0, 0)$.

Using both the projection W and the centers μ learnt during the training step, a new query x in the test set (a dimension d row vector) is classified according to the following rule: it belongs to the cluster number j^* if and only if

$$j^* \in \arg \min_{j=1,\ldots,k} \|\mu_j - xW\|_1. \tag{3}$$

2.1.2 Primal-dual scheme, constrained formulation

To handle features with a high correlation, we consider a convex constrained supervised classification problem. However the drawback of the term $\|Y\mu - XW\|_1$ is that it enforces equality of the two matrices out of a sparse set: hence it tunes the parameters to enforce a perfect matching of the training data. We replace the 1-norm with the robust "Huber function" [13]. If $h_\delta(t) = t^2/(2\delta)$ for $|t| \leq \delta$ and $|t| - \delta/2$ for $|t| \geq \delta$.

We obtain the following criterion

$$\min_{(W,\mu)} h_\delta(Y\mu - XW) + \frac{\rho}{2}\|I_k - \mu\|_F^2 \text{ s.t.} \quad \|W\|_1 \leq \eta. \tag{4}$$

We can tune a primal-dual method to solve this problem with Algorithm 1 (See [13] and [27] for details)

Algorithm 1 Primal-dual algorithm, constrained case—$proj(V, \eta)$ is the projection on the ℓ_1 ball of radius η

Input: $X, Y, N, \sigma, \tau, \tau_\mu, \eta, \delta, \rho, \mu_0, W_0, Z_0$
for $n = 1, \ldots, N$ **do**
  $W_{\text{old}} := W$
  $\mu_{\text{old}} := \mu$
  $W := W + \tau \cdot (X^T Z)$
  $W := proj(W, \eta)$
  $\mu := \frac{1}{1+\tau_\mu \cdot \rho}(\mu_{\text{old}} + \rho \cdot \tau_\mu I_k - \tau_\mu \cdot (Y^T Z))$
  $Z := \frac{1}{1+\sigma \cdot \delta}(Z + \sigma \cdot (Y(2\mu - \mu_{\text{old}}) - X(2W - W_{\text{old}})))$
  $Z := \max(-1, \min(1, Z))$
end for
Output: W, μ

2.1.3 Classification with rejection using a confidence Score for the Prediction (CSP)

False positive (FP) and false negative (FN) results are an important issue for diagnostic tools in medicine. One way to diminish the number of FP and FN results is to use classification with rejection [19, 28] for which classifiers are allowed to report "I don't know". This type of classification enables the incorporation of doubt in the results if the observation x is too hard to classify. Here, we propose to use a confidence score for the prediction (CSP) to devise a classifier with rejection.

In our analysis we only had two clusters with centers μ_1 and μ_2 Lets recall that the predicted label j^* of a sample x is given by

$$j^* \in \arg\min_{j=1,\ldots,2} \|\mu_j - xW\|_1. \tag{5}$$

We can compute the distances of sample x to the two centroids, respectively. $d_1 = \|\mu_1 - xW\|_1$ and $d_2 = \|\mu_2 - xW\|_1$ and we propose a confidence indicator for sample x as follows :

$$\rho(x) = \frac{d_1 - d_2}{d_1 + d_2} \tag{6}$$

Thus, the CSP $\rho(x)$ is a value ranging from -1 to 1. The closer the CSP $\rho(x)$ is to +1 or -1 depending on the predicted class, the higher the confidence for the prediction will be.

Thus if ϵ is a given threshold parameter, we can perform classification with rejection by rejecting binary classification for samples with an absolute value of CSP $\rho(x)$ under this threshold. The labels will then be

predicted as follows :

$$Label = \begin{cases} -1 & \text{if } \rho(x) < -\epsilon \\ 0 & \text{if } -\epsilon < \rho(x) < \epsilon \\ 1 & \text{if } \rho(x) > \epsilon \end{cases} \tag{7}$$

We can then study the False Discovery Rate (FDR) $FDR = FP + FN$ as a function of parameter ϵ.

2.2 Availability of the method

We implemented PD-CR in python. Functions and scripts are freely available at https://github.com/tirolab/PD-CR.

3 Comparison to PLS-DA, Random Forests and SVM using 2 datasets

To compare PD-CR to the standard PLS-DA, Random Forests and SVM classification methods in terms of accuracy and feature selection, we tested the four methods on two metabolomic datasets named "BRAIN" and "LUNG". Accuracies and feature selection for each method were obtained using 4 fold-cross validation with varying random seeds. We also provide the results with a a new version of PD-CR minimizing the $\ell 2$ norm PD-CR $\ell 2$ (See Algorithm 6 $https://arxiv.org/pdf/1902.01600.pdf$).

3.1 LUNG dataset

The LUNG dataset was provided by Mathe *et al.* This dataset includes metabolomics data concerning urine samples from 469 Non-Small Cell Lung Cancer (NSCLC) patients prior to treatment and 536 controls collected from 1998 to 2007 in seven hospitals and in the Department of Motor Vehicles (DMV) from the greater Baltimore, Maryland area. Urine samples were analyzed using an unbiased metabolomics LC-MS/MS approach. This dataset is available from the MetaboLights database (study identifier MTBLS28)

Mathe *et al.* used Random Forests to classify patients as lung cancer patients or controls[10]. The aim was to create a new screening test for lung cancer, based on metabolomics data from urine. Lung cancer is one of the most common cancers and it is well established that early diagnosis is essential for treatment. An efficient screening method based on urinary metabolomics would be of great benefit.

3.2 BRAIN dataset

The BRAIN dataset was obtained from a metabolomic study performed by our biological team (TIRO) on frozen samples of glial tumors. The samples were provided by the university hospitals of Nice and Montpellier (France). Metabolite extracts were prepared and analyzed in the TIRO laboratory (Nice, France). With this dataset, the goal was to create a model that accurately discriminated between mutated isocitrate dehydrogenase (IDH) and IDH wild-type glial tumors. This mutation is a key component of the World Health Organization classification of glial tumors [29]. The mutational status is usually assessed by IDH1 (R132H)-specific (H09) immunohistochemistry. Yet this technique can lead to False-Negative results [30], which can only be identified by sequencing. An accurate metabolomic based test, able to assess the IDH mutational status, could be a promising solution to this problem.

These samples were retrospectively collected from two declared biobanks from the Central Pathology Laboratory of the Hospital of Nice and from the Center of Biological Resources of Montpellier (Plateforme CRB-CHUM). Consent or non-opposition was verified for every participant. For every participant, the IDH mutational status was assessed using immunohistochemistry and pyrosequencing for immunonegative cases.

Samples of brain tumors were analyzed using Liquid Chromatography coupled to tandem Mass Spectrometry (LC-MS/MS) in an unbiased metabolomics approach, as performed in a previous metabolomics study [**?**].

The details of the analysis are available in supplementary material.

3.3 Data Filtering and Pre-processing

Our laboratory performed the LC-MS/MS analysis for the BRAIN dataset. Therefore, we could apply different levels of filtering on this dataset. After processing of the raw data using MZmine 2.39 software, two types of filtering were applied to the BRAIN dataset, minimal and maximal filtering. The minimal filtering only removed metabolites for which a spike was detected in less than 10 percent of the samples. The maximal filtering removed all unidentified metabolites as well as metabolites that did not have an isotopic pattern. This filtering method is frequently used for metabolomic studies and diminishes the number of noisy features in the dataset. Furthermore, it diminishes the time necessary for data processing because it diminishes the data volume. Unfortunately, any filtering will necessarily come with a high risk of removing some relevant features which is also the case with this filtering method. Using the two BRAIN datasets, we aimed to assess how the filtering affected the results of the different classification methods. The LUNG dataset was used as it was published, without additional normalization or filtering.

3.4 Availability of the data

The datasets are freely available at https://github.com/tirolab/PD-CR.

3.5 Comparison to other methods :

Before comparison, the data were pre-processed as follows:
i) Log-transformation for the following benefits: Reducing heteroscedasticity and thus the bias on regression and transforming multiplicative noise into additive noise,
ii) Mean centering and scaling [31].

PD-CR [13] was compared to PLS-DA[32], Random Forests (with 100 and 400 trees)[9] and SVM using the sklearn python package.

Additionally, we evaluated the impact of the use of the Huber loss in PD-CR compared to the use of the $\ell 2$ loss.

Parameters σ, τ, δ and ρ were set according to results obtained using various datasets in an initial step [13] and were not further tuned. Parameter η, which affects the feature selection step was manually tuned to fit the number of features in the datasets and to maximize accuracy after cross validation.

We computed the accuracy of the 4 classification methods for the two metabolomics datasets using 4-fold cross-validation (Script "PD-CR vs PLS-DA, RF and SVM" on https://github.com/tirolab/PD-CR). The selected metabolites were analyzed and compared between methods for the metabolomics datasets.

For PD-CR, we plotted the histograms of the CSP $\rho(x)$ and the probability distribution function (PDF) as well as the False Discovery Rate (FDR =(FP+FN)/total) and the rate of rejected samples (RRS = rejected samples/total samples) depending on epsilon (the CSP threshold) (Script "rhoComputing" on https://github.com/tirolab/PD-CR).

4 Results

The characteristics of the two metabolomics datasets are presented in Table 1 .

The LUNG dataset included a large number of patients (a little over 1,000) with an equivalent number of features (a little under 3,000) and the BRAIN dataset included a smaller number of patients (88) with a much higher number of features. While obtaining metabolomics data concerning as many patients as there are in the LUNG dataset is remarkable, the number of patients in the BRAIN dataset is closer to the number of patients in most metabolomics studies.

Dataset	No. of Samples	No. of features	Sample type
LUNG	1005	2944	Urine
BRAIN	88	25,286	Glial tumor tissue

Table 1: Overview of the datasets.

4.1 LUNG:

LUNG	PD-CR	PD-CR ℓ_2	PLS-DA	RF (100 trees)	RF (400 trees)	SVM
Accuracy %	79.44	78.3	76.56	71.31	72.44	76.25
AUC	79.97	-	74.05	73.38	74.50	76.64
Time (s)	0.11	0.11	0.09	0.89	3.47	85.6

Table 2: **LUNG** dataset: Mean accuracy using 3 seeds and 4-fold cross validation: comparison with PLS-DA , Random forest and Best SVM

As shown in Table 2, PD-CR outperformed PD-CR ℓ_2, PLS-DA, Random Forests (400 trees) and SVM by 1.1%, 2.8%, 7% and 3.1% respectively.

Even though an accuracy of 79.44% may be high enough to consider using our PD-CR method and urinary metabolomics for the screening of lung cancer, Figure 1 shows that the accuracy may be even higher if the CSP is taken into account and if it is used to perform classification with rejection. Indeed, in Figure 1 the top left shows the histogram of the CSP and the top right the kernel probability distribution function (PDF). We can see that healthy controls and cancer patients are predicted with an equally high confidence. On the bottom left the False Discovery Rate ($FDR = (FP+FN)/total\ samples$) decreases as the confidence score threshold increases, but as shown in the bottom right, the rate of rejected samples ($RRS = rejected\ samples/total\ samples$) increases.

As shown in Table 3, PD-CR selected "MZ 264.1215224" for a molecular ion at m/z 264.1215224 and "MZ 308.0984878" for a molecular ion at m/z 308.0984878 as the top two features.

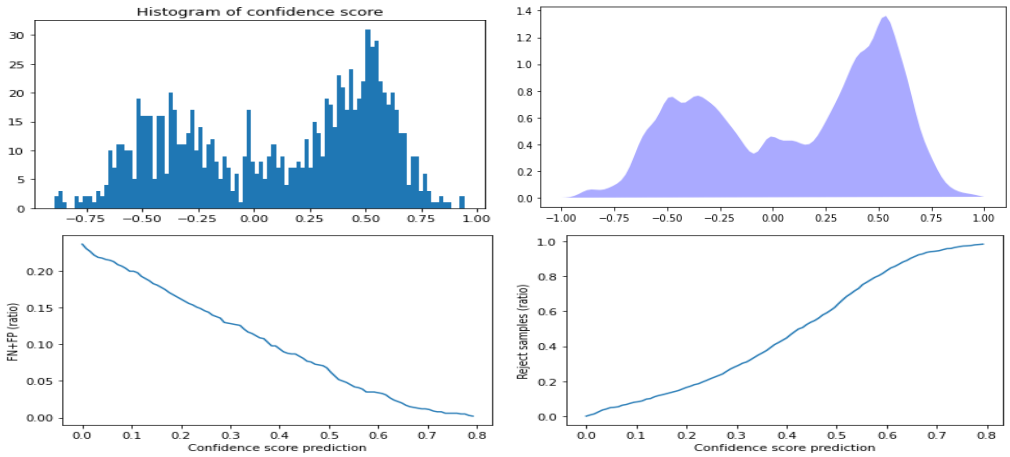

Figure 1: *Distribution of the Confidence Score for the Prediction (CSP) on the Lung dataset and impact of using CSP for classification with rejection on the false discovery rate (FDR). From Left to right and top to bottom : Histogram of the CSP, Kernel density estimation; FDR as a function of CSP after classification with rejection, rate of rejected samples as a function of CSP after classification with rejection. As expected for a pertinent confidence score, the FDR diminishes when using a higher CSP threshold for classification with rejection.*

RF	PLS-DA	PD-CR	SVM
MZ 264.1215224	MZ 264.1215224	MZ 264.1215224	MZ 264.1215224
MZ 656.2017529	MZ 126.9069343	MZ 308.0984878	MZ 308.0984878
MZ 441.1613664	MZ 170.0605916	MZ 126.9069343	MZ 247.0970455
MZ 584.2670695	MZ 613.3595637	MZ 613.3595637	MZ 613.3595637
MZ 247.0970455	MZ 243.1004849	MZ 243.1004849	MZ 615.0353192
MZ 486.2571336	MZ 486.2571336	MZ 247.0970455	MZ 372.9232556
MZ 308.0984878	MZ 308.0984878	MZ 332.0963401	MZ 441.1613664
MZ 204.1345526	MZ 561.3432022	MZ 441.1613664	MZ 370.0525988
MZ 247.1384435	MZ 94.06574518	MZ 94.06574518	MZ 423.0084949
MZ 447.10803	MZ 269.1280232	MZ 561.3432022	MZ 332.0963401

Table 3: Top 10 features selected by Random Forests, PLS-DA, PD-CR and SVM in the LUNG dataset

These features "MZ 264.1215224" and "MZ 308.0984878" most likely correspond to creatine riboside (expected m/z value in the positive mode: 264.1190; mass error: 10 ppm) and N-acetylneuraminic acid (expected m/z value in the negative mode : 308.0987; mass error: 1 ppm), respectively. These two metabolites were described by Mathé *et al.* [10] as the two most important metabolites to discriminate between lung cancer patients and healthy individuals using Random Forests on metabolomic data from urine samples. Indeed, these two metabolites were significantly higher in the urines of lung cancer patients, as shown in Figure 2.

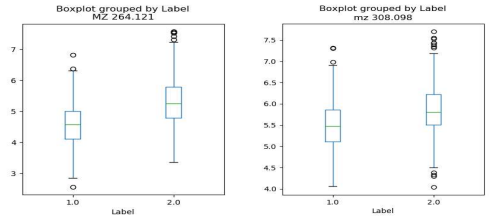

Figure 2: Boxplots concerning relative abundances of features MZ 264.1215224 and MZ 308.0984878 of the LUNG dataset, most likely corresponding to creatine riboside and N-acetylneuraminic acid respectively. Fold changes : 2.57 and 1.43 respectively. Label 1 indicates urine samples of patients without lung cancer. Label 2 indicates urine samples of patients with lung cancer.

4.2 BRAIN:

4.2.1 Minimally filtered dataset :

BRAIN	PD-CR	PD-CR ℓ_2	PLS-DA	RF (100 trees)	RF (400 trees)	SVM
Accuracy %	92.04	90.9	84.09	88.63	89.39	87.78
AUC	92.08	-	84.33	88.70	89.02	88.53

Table 4: **BRAIN** dataset Accuracy using 3 seeds and 4-fold cross validation: comparison with PLS-DA, Random Forest and best SVM.

As shown in Table 4, PD-CR outperformed PD-CR ℓ_2, PLS-DA, Random Forests (400 trees) and SVM by 1.1%, 7.7%, 2.7% and 4.3%, respectively for the BRAIN dataset. For this high dimensional dataset, the number of features (25,286) significantly exceeded the number of samples (88) giving a significant drop in the PLS-DA accuracy.

Furthermore, as shown in Figure 3 the accuracy obtained with PD-CR could be further improved by using the CSP to perform classification with rejection. Indeed, most of the samples were classified with a high CSP and if we apply a CSP threshold ϵ of 0.45, the FDR drops to 0 while only rejecting 10% of the samples. This shows that all the miss-classified samples had a low CSP.

As shown in Table 5, most of the top features selected with the 3 methods correspond to different isotopes and adducts of 2-hydroxyglutarate. Indeed, POS_MZ131.0342, POS_MZ132.0375 and POS_MZ133.0384 all

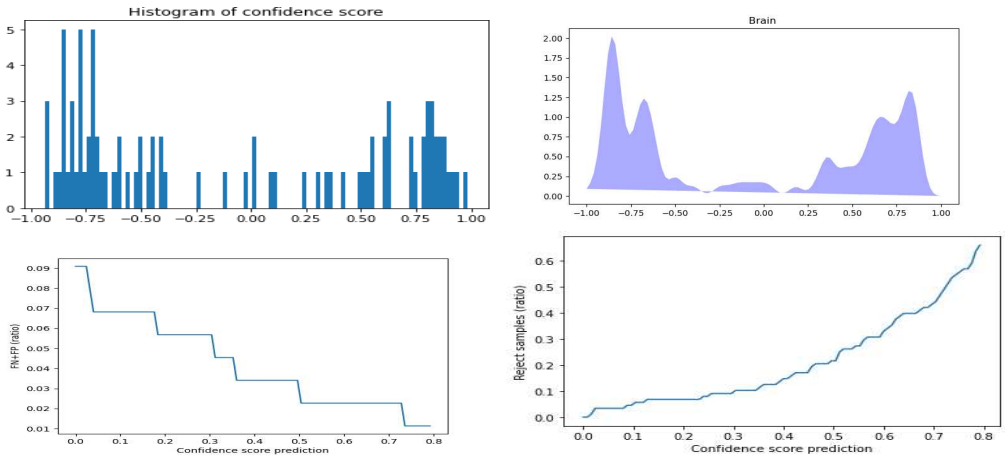

Figure 3: Distribution of the Confidence Score for the Prediction (CSP) on the BRAIN dataset and impact of using CSP for classification with rejection on the false discovery rate (FDR). From left to right and top to bottom : Histogram of the CSP, Kernel density estimation; FDR as a function of CSP after classification with rejection, rate of rejected samples as a function of CSP after classification with rejection.As expected for a pertinent confidence score, the FDR diminishes when using a higher CSP threshold for classification with rejection.

Random Forests	**PLS-DA**	**PD-CR**	**SVM**
NEG_MZ147.0867	POS_MZ131.0342	POS_MZ131.0342	$POS_M Z131.0342$
POS_MZ133.0384	POS_MZ132.0375	POS_MZ132.0375	$POS_M Z132.0375$
POS_MZ166.0713	POS_MZ166.0713	POS_MZ243.9903	$POS_M Z166.0713$
POS_MZ228.0182	NEG_MZ147.0288	POS_MZ166.0712	$NEG_M Z147.0288$
POS_MZ132.5234	NEG_MZ148.0321	NEG_MZ147.0288	$NEG_M Z148.0321$
POS_MZ173.0306	NEG_MZ149.0329	NEG_MZ148.0321	$POS_M Z171.0265$
POS_MZ219.0082	POS_MZ171.0265	POS_MZ123.5181	$OS_M Z132.0375$
NEG_MZ215.0168	POS_MZ132.0375	POS_MZ171.0265	$POS_M Z247.9616$
POS_MZ171.0265	POS_MZ243.9903	NEG_MZ149.0329	$POS_M Z243.9903$
POS_MZ319.0510	POS_MZ123.5181	POS_MZ133.0384	$NEG_M Z149.0329$

Table 5: Top 10 features selected by Random Forests, PLS-DA, PD-CR and SVM on the BRAIN dataset with 25286 features

correspond to the [M+H-H2O adduct]+ of 2-hydroxyglutarate with C12, and two C13 isotopes respectively. NEG_MZ147.0288, NEG_MZ148.0321 and NEG_MZ149.0329 correspond to the [M-H]- adduct with C12, and two C13 isotopes respectively. POS_MZ166.0713 corresponds to a [M+NH4]+ adduct. POS_MZ171.02645 corresponds to the [M+Na]+ adduct. POS_MZ243.9903 had the same retention time and chromatographic profile as POS_MZ131.0342, suggesting that it was an unknown fragment or adduct of 2-hydroxyglutarate. 2-Hydroxyglutarate is a well-known oncometabolite produced in high quantities by mutated IDH1/2 in gliomas [33]. It is therefore expected that this compound will have a high weight when classifying mutated vs wild-type gliomas as it should be significantly increased in IDH mutated gliomas (as shown in figure 4).

Here all four methods selected this important feature among a high dimensional dataset (25,287 features in this case). Adducts and isotopes of 2-hydroxyglutarate with low levels are top selected features using PD-CR indicating that our method is a very sensitive way to identify significant molecules. This result on the minimally filtered dataset also suggest that PC-CR avoids overfiting as no unexpected feature was selected.

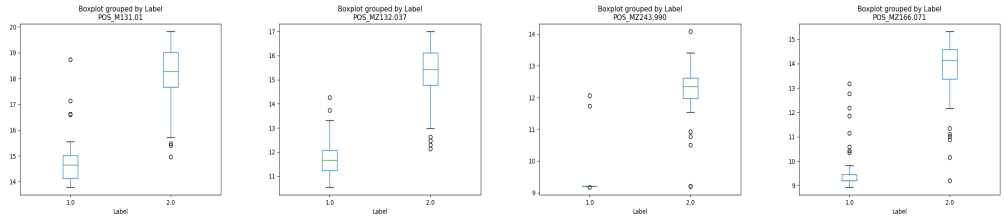

Figure 4: Boxplots concerning relative abundances of features POS_131.0342, POS_132.0375 POS_243.9903 and POS_166.0712 of the BRAIN dataset, most likely corresponding to different adducts of 2-Hydroxyglutarate. Fold changes : 32.9, 35.6, 14.6 and 33.7 respectively. Label 1 : samples of tumors with wild type IDH, Label 2 : samples of tumors with mutated IDH.

4.2.2 Comparison to the highly filtered dataset

	PD-CR	PD-CR ℓ_2	PLS-DA	Random Forests	SVM
Accuracy %	94.31	92.8	93.18	92.04	89.20

Table 6: Mean accuracy using 4-fold cross validation with 3 different seeds: comparison of methods on the **BRAIN highly filtered** data set

As shown in Table 6 the accuracies of the different methods were equivalent and very high when using the highly filtered version of the BRAIN dataset (accuracy being a little lower with SVM).

When PD-CR was used on the highly filtered BRAIN dataset, it lead to similar results as with PD-CR using an $\ell 2$ loss, PLS-DA, Random Forests and SVM. In contrast, it outperformed these methods when using the minimally filtered dataset. In this case, as shown in Table 7 more features were selected. When using the BRAIN dataset for the IDH-mutated vs wild-type classes, most of these additional features were adducts of 2-hydroxyglutarate and are therefore known to be biologically relevant. The additional features that are not

adducts of 2-hydroxyglutarate will be investigated in a future study.

Identified (495 features)	Large (25287 features)
POS_M131.0342	**POS_MZ131.0342**
NEG_M147.02882	POS_MZ132.0375
POS_M85.0291	POS_MZ243.9903
POS_M149.0450	POS_MZ166.0713
NEG_M112.0220	**NEG_MZ147.0288**
POS_M154.0864	NEG_MZ148.0320
NEG_M171.0847	POS_MZ123.518
NEG_M320.0627	POS_MZ171.0265
POS_M113.0350	NEG_MZ149.0329
POS_M147.1170	POS_MZ133.0384

Table 7: Top 10 features selected by PD-CR in the highly and minimally filtered versions of the **BRAIN** dataset

5 Discussion

Machine learning methods are of particular interest for metabolomics studies and are being used increasingly for other omics studies. Herein we introduce a new primal-dual method for supervised classification and feature selection. To our knowledge, a primal-dual method had never been used in this way. We compare this method to three of the most frequently used methods: PLS-DA, Random Forests and SVM, on two metabolomics datasets. Metabolomics datasets tend to be sparse datasets including highly correlated features. PD-CR is particularly suited for this data structure. Hence, for metabolomics, PD-CR appears to be more accurate than the three other methods while selecting biologically relevant features and providing a confidence score for each prediction. An important upside associated with the inclusion of a confidence score for each prediction is that it enables classification with rejection.

We believe that this confidence score is of great value, particularly for applications in medicine. Metabolomics approaches are of particular interest for medical applications. Indeed, they could be used in routine clinical practice as they are relatively inexpensive and can be performed rapidly compared to proteomics, transcriptomics or genomics analyses. More and more studies suggest that metabolomics associated to classification methods are very promising tools for individual personalized medicine[34, 10]. To use metabolomics in routine clinical practice it is paramount to obtain robust, rapid and trustworthy predictions. The confidence score provided with PD-CR adds considerable value to the prediction as it includes a metric that is implicitly used by every physician when they make a medical decision: the probability to make the wrong choice. So far, one of the main obstacles to the use of machine learning in medicine resides in the fact that it is harder to trust

the decision of a machine learning method than that of a physician when it comes to health issues. We believe that providing a confidence score associated to the decision would make these new tools more convincing if used in routine clinical practice. Furthermore, this confidence score can be used to perform classification with rejection and reduce the false discovery rate.

Furthermore, this confidence score could be extended to more than 2 classes as follows : We can compute the distances of sample x to all the centroids, respectively. $d_1 = \|\mu_i - xW\|_1$ and we propose a confidence indicator for sample x as follows :

$$\rho(x) = 1 - k\frac{Min(d_1, d_2..., d_k)}{d_1 + d_2 + ...d_k} \tag{8}$$

Thus, the CSP $\rho(x)$ is a value ranging from 0 to 1. The closer the CSP $\rho(x)$ is to +1 for a predicted class, the higher the confidence will be.

We have shown that PD-CR outperformed the common PLS-DA, Random Forests and SVM methods on both LUNG and BRAIN datasets. We believe that this is partly due to the fact that PD-CR uses a Huber loss. Indeed, the use of the Huber loss with PD-CR leads to a better accuracy than the use of a common $\ell 1$ or $\ell 2$ loss [13]. Note that the l1 loss is not derivable in zero. Moreover the drawback of the term $\|Y\mu - XW\|_1$ of the l1 loss is that it enforces equality of the two matrices out of a sparse set. Moreover the use of the Huber loss reduces the impact of the presence of outliers in the training set, and therefore leads to a better accuracy than the $\ell 2$ loss, as shown in table 2 and table 4.

Furthermore we show in table 2 and table 4 that using PD-CR with an $\ell 2$ loss provides better results than PLSDA which uses the same $\ell 2$ loss. This is probably due to the fact that PLS-DA does not perform feature selection and is known to be prone to overfitting [5].

Moreover, when comparing methods with the minimally filtered and the more filtered versions of the BRAIN dataset, all methods suffered a decrease in accuracy with the minimally filtered dataset (PD-CR keeping the higher accuracy). However the results obtained using the PLS-DA method appeared to be more impacted than those of the Random Forests, SVM and PD-CR. Indeed, the accuracy of PLS-DA significantly decreased when the less filtered dataset was used dropping from 93.18% to 84.09%, compared to a mild decrease in accuracy for the other methods. This can also be explained by the fact that PLS-DA does not perform feature selection and is known to be prone to overfitting [5]. For this reason, several strategies are commonly used to reduce the number of features in metabolomics datasets. Features can be filtered according to the number of

detected peaks in all samples, the correct identification of the compound (using the most common adduct) or the presence of isotopes. Working with filtered data has some advantages, including the fact that it appears more biologically relevant to work on less noisy and more reliable data. However, filtering also has some important drawbacks, the most important being the high risk of removing interesting metabolites from the dataset. In the case of the BRAIN dataset, 2-Hydroxyglutarate is a well known metabolite associated to IDH mutation. However, in many metabolomic studies, the goal is to discover potentially unidentified metabolites associated to particular conditions which can only be achieved by including unidentified metabolites. As shown in this work, PD-CR can be applied to both minimally filtered and highly filtered metabolomics datasets.

As it has been previously reported, when designing prediction models, some methods may lead to a more accurate model for a specific dataset while others may be more adapted with other datasets [35]. Indeed, even though we can discuss which machine learning method is the best, most often, researchers try out several machine learning methods on their metabolomics datasets and report the results of the most accurate one. This process has even been automated by some authors [36]. PD-CR is an advanced method, based on recent development in convex optimization and we believe it should be considered by researchers when designing prediction models for metabolomics studies.

Much like the commonly used methods PLS-DA, Random Forests and SVMs, available with [37], our python implementation of PD-CR only requires the tuning of one parameter : η. This makes the use of PD-CR quite simple, even for non machine learning experts, much like PLS-DA. Note that the tuning of the η parameter must be done carefully since it modifies feature selection.

When comparing misclassified patients between methods in an additional analysis, it appeared that in the minimally filtered BRAIN dataset 16/88 tumors were misclassified with at least one method. 2 tumors were misclassified with all methods, 6 with two or three methods and 8 with only one method (3 were misclassfified only with PLS-DA, 4 with Random Forests, 1 with SVM and none with PD-CR). In the LUNG dataset 702/1005 patients were misclassified with at least one method. 68 patients were misclassified with all methods, 240 with two or three methods and 394 with only one method (15 were misclassfified only with PLS-DA, 63 with Random Forests, 305 with SVM and 11 with PD-CR). It therefore appears that PD-CR is the method with the smallest number of false discoveries.

While prior metabolomic studies did not necessarily focus on validating which features the prediction models relied on, it is now admitted that to be trustworthy a model must be based on biologically relevant features and must therefore be interpretable [38]. Indeed, interpretability of machine learning methods [39] is crucial to assess if selected features are biologically relevant. PD-CR offers a straightforward, reliable metric based on the weights of each feature in the model (matrix W).

Conversely, non-linear methods such as Random Forests or non-linear SVM and the linear methods PLS-DA and linear SVM are usually associated to method-specific metrics which makes it difficult to compare features between methods. For Random Forests, the Mean Decrease Impurity (MDI) is usually the default metric for variable importance [40]. It is computed as a mean of the individual trees' improvement in the splitting criterion produced by each variable. For PLS-DA, the Variable Importance for the Projection (VIP) score is often used. The VIP score is computed by summing the contributions VIN (variable influence) over all model dimensions. For a given PLS dimension a, $(VIN)^2_{ak}$ is a function of the squared PLS weight w^2_{ak} [41].

While these metrics offer some insight into the importance of each metabolite in the model these are indirect metrics whereas the weights provided with PD-CR represent the direct quantitative measure of the importance of each feature in the model, very close to regression parameters and can thus directly be used to classify a new sample.

Furthermore, relevant feature selection is necessary for a correct understanding of the biological mechanisms underlying classification. It is well established that when expressed, mutant IDH 1/2 reduces 2-oxo-glutarate to 2-hydroxyglutarate [42]. It was therefore expected for 2-hydroxyglutarate to be a feature of importance as was the case when using PD-CR on the BRAIN dataset for the classification of IDH-mutated vs wild-type gliomas. As the biologically relevant features are known in advance, the BRAIN dataset is a good testing set for this new method. Furthermore, as we described, the features selected with PD-CR in the LUNG dataset are identical to the ones described by Mathé *et al.* in their original study, which also validates the accurate feature selection performed by PD-CR.

6 Conclusion

Herein we propose a recently introduced primal-dual method (PD-CR) for feature selection and classification with rejection . To our knowledge, the primal-dual method has never been used in such fashion. PD-CR includes a sparse regularization factor which is particularly appropriate for high dimensional sparse datasets

such as metabolomics datasets.

We highlight the two main results. First, PD-CR is more accurate than PLS-DA, Random Forests and SVM and leads to the selection of biologically relevant features. Second, our method provides a confidence score for each prediction and allows classification with rejection, which can help reduce false discovery rates.

References

[1] Johnson, C.H., Ivanisevic, J., Siuzdak, G.: Metabolomics: beyond biomarkers and towards mechanisms. Nature reviews. Molecular cell biology **17**(7), 451–459 (2016). doi:10.1038/nrm.2016.25. Accessed 2018-06-15

[2] Kell, D.B.: Metabolomics and systems biology: making sense of the soup. Current Opinion in Microbiology **7**(3), 296–307 (2004). doi:10.1016/j.mib.2004.04.012. Accessed 2020-04-05

[3] Barker, M., Rayens, W.: Partial least squares for discrimination. Journal of Chemometrics **17**(3), 166–173 (2003)

[4] Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., Goodacre, R.: A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. Analytica Chimica Acta **879**, 10–23 (2015). doi:10.1016/j.aca.2015.02.012. Accessed 2020-04-05

[5] Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J.J., van Duijnhoven, J.P.M., van Dorsten, F.A.: Assessment of PLSDA cross validation. Metabolomics **4**(1), 81–89 (2008). doi:10.1007/s11306-007-0099-6. Accessed 2020-04-06

[6] Brereton, R.G.: Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. TrAC Trends in Analytical Chemistry **25**(11), 1103–1111 (2006). doi:10.1016/j.trac.2006.10.005. Accessed 2020-04-06

[7] Broadhurst, D.I., Kell, D.B.: Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics **2**(4), 171–196 (2006). doi:10.1007/s11306-006-0037-z. Accessed 2020-04-06

[8] Bartel, J., Krumsiek, J., Theis, F.J.: STATISTICAL METHODS FOR THE ANALYSIS OF HIGH-THROUGHPUT METABOLOMICS DATA. Computational and Structural Biotechnology Journal **4**(5), 201301009 (2013). doi:10.5936/csbj.201301009. Accessed 2020-04-06

[9] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)

[10] Mathé *et al*, E.: Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. Cancer research **74**(12), 3259–3270 (2014)

[11] Heinemann, J., Mazurie, A., Tokmina-Lukaszewska, M., Beilman, G.J., Bothner, B.: Application of support vector machines to metabolomics experiments with limited replicates. Metabolomics **10**(6), 1121–1128 (2014). doi:10.1007/s11306-014-0651-0. Accessed 2021-05-27

[12] Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning **46**(1-3), 389–422 (2002)

[13] Barlaud, M., Chambolle, A., Caillau, J.-B.: Classification and feature selection using a primal-dual method and projection on structured constraints. International Conference on Pattern Recognition, Milan, 6538–6545 (2020)

[14] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288 (1996)

[15] Jacob, L., Obozinski, G., Vert, J.-P.: Group lasso with overlap and graph lasso. In: Proceedings of the 26th International Conference on Machine Learning (ICML-09), pp. 353–360 (2009)

[16] Ding, C., Li, T.: Adaptive dimension reduction using discriminant analysis and k-means clustering. In: Proceedings of the 24th International Conference on Machine Learning, pp. 521–528 (2007)

[17] Combettes, J.-C. P.L.and Pesquet: A douglas-rachford splitting approach to nonsmooth convex variational signal recovery. IEEE J. Selected Topics Signal Process., 564–574 (2007)

[18] Barlaud, M., Belhajali, W., Combettes, P.L., Fillatre, L.: Classification and regression using an outer approximation projection-gradient method. IEEE Transactions on Signal Processing **65**(17), 4635–4644 (2017)

[19] Freund, M.Y. Y., Schapire, R.E.: Generalization bounds for averaged classifiers. Annals of Statistics **32**(4), 1698–1722 (2004)

[20] Nock, R., BelHajAli, W., Dambrosio, R., Nielsen, F., Barlaud, M.: Gentle nearest neighbors boosting over proper scoring rules. vol. 37, pp. 80–93. IEEE (2015)

[21] Hastie, T., Rosset, S., Tibshirani, R., Zhu, J.: The entire regularization path for the support vector machine. Journal of Machine Learning Research **5**, 1391–1415 (2004)

[22] Friedman, J., Hastie, T., Tibshirani, R.: Regularization path for generalized linear models via coordinate descent. Journal of Statistical Software **33**, 1–122 (2010)

[23] Hastie, T., Tibshirani, R., Wainwright, M.: Statistcal learning with sparsity: The lasso and generalizations. CRC Press (2015)

[24] Li, J., Cheng, K., Wang, S., Morstatter, F., P. Trevino, R., Tang, J., Liu, H.: Feature selection: A data perspective. ACM Computing Surveys **50** (2016)

[25] Ali, A., Tibshirani, R.: The generalized lasso problem and uniqueness. Electronic Journal of Statistics **13**(2), 2307–2347 (2019)

[26] McQueen, J.-B.: Some methods for classification and analysis of multivariate observations. Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability (1967)

[27] Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal-dual algorithm. Math. Program. **159**(1-2, Ser. A), 253–287 (2016)

[28] Ni, C., Charoenphakdee, N., Honda, J., Sugiyama, M.: On the Calibration of Multiclass Classification with Rejection (2019). 1901.10655

[29] Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., Ellison, D.W.: The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathologica **131**(6), 803–820 (2016). doi:10.1007/s00401-016-1545-1. Accessed 2020-04-09

[30] Yoshida, A., Satomi, K., Ohno, M., Matsushita, Y., Takahashi, M., Miyakita, Y., Hiraoka, N., Narita, Y., Ichimura, K.: Frequent false-negative immunohistochemical staining with IDH1 (R132H)-specific H09 antibody on frozen section control slides: a potential pitfall in glioma diagnosis. Histopathology **74**(2), 350–354 (2019). doi:10.1111/his.13756

[31] van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J.: Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC genomics **7** (2006)

[32] Wold, S., Sjostrom, M., Eriksson, L.: Pls-regression: a basic tool of chemometrics. Elsevier volume 58, issue 2, 109–130 (2001)

[33] Dang, L., White, D.W., Gross, S., Bennett, B.D., Bittinger, M.A., Driggers, E.M., Fantin, V.R., Jang, H.G., Jin, S., Keenan, M.C., Marks, K.M., Prins, R.M., Ward, P.S., Yen, K.E., Liau, L.M., Rabinowitz,

J.D., Cantley, L.C., Thompson, C.B., Vander Heiden, M.G., Su, S.M.: Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature **462**(7274), 739–744 (2009). doi:10.1038/nature08617

[34] Jing, L., Guigonis, J.-M., Borchiellini, D., Durand, M., Pourcher, T., Ambrosetti, D.: LC-MS based metabolomic profiling for renal cell carcinoma histologic subtypes. Scientific Reports **9**(1), 1–10 (2019)

[35] Madsen, R., Lundstedt, T., Trygg, J.: Chemometrics in metabolomics—A review in human disease diagnosis. Analytica Chimica Acta **659**(1), 23–33 (2010). doi:10.1016/j.aca.2009.11.042. Accessed 2020-05-11

[36] Leclercq, M., Vittrant, B., Martin-Magniette, M.L., Scott Boyer, M.P., Perin, O., Bergeron, A., Fradet, Y., Droit, A.: Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. Frontiers in Genetics **10** (2019). doi:10.3389/fgene.2019.00452. Accessed 2020-05-08

[37] Xia, J., Psychogios, N., Young, N., Wishart, D.S.: MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Research **37**(suppl$_2$), 652 – −660(2009).*doi* : 10.1093/nar/gkp356.https://academic.oup.com/nar/article-pdf/37/suppl_2/W652/3933058/gkp356.pdf

[38] Zhang, A., Sun, H., Yan, G., Wang, P., Wang, X.: Mass spectrometry-based metabolomics: applications to biomarker and metabolic pathway research. Biomedical Chromatography **30**(1), 7–12 (2016). doi:10.1002/bmc.3453. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bmc.3453. Accessed 2020-05-13

[39] Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning (2017). 1702.08608

[40] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, : Scikit-learn: Machine Learning in Python. arXiv:1201.0490 [cs] (2018). arXiv: 1201.0490. Accessed 2020-05-11

[41] Chong, I.-G., Jun, C.-H.: Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems **78**(1), 103–112 (2005). doi:10.1016/j.chemolab.2004.12.011. Accessed 2020-05-13

[42] Losman, J.-A., Kaelin, W.G.: What a difference a hydroxyl makes: mutant IDH, (R)-2-hydroxyglutarate, and cancer. Genes & Development **27**(8), 836–852 (2013). doi:10.1101/gad.217406.113. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold

Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Accessed 2020-05-19

Article 2 : Learning a confidence score and the latent space of a new supervised autoencoder for diagnosis and prognosis in clinical metabolomic studies

David Chardin, Cyprien Gille, Thierry Pourcher, Olivier Humbert, Michel Barlaud

BMC Bioinformatics. 2022 Sep 1;23(1):361. doi: 10.1186/s12859-022-04900-x.

Learning a confidence score and the latent space of a new supervised autoencoder for diagnosis and prognosis in clinical metabolomic studies.

David Chardin, Olivier Humbert, Caroline Bailleux, Fanny Burel-Vandenbos, Valerie Rigau, Thierry Pourcher, Michel Barlaud

01/09/2022

Abstract

Background: Presently, there is a wide variety of classification methods and deep neural network approaches in bioinformatics. Deep neural networks have proven their effectiveness for classification tasks, and have outperformed classical methods, but they suffer from a lack of interpretability. Therefore, these innovative methods are not appropriate for decision support systems in healthcare. Indeed, to allow clinicians to make informed and well thought out decisions, the algorithm should provide the main pieces of information used to compute the predicted diagnosis and/or prognosis, as well as a confidence score for this prediction.

Methods:Herein, we used a new supervised autoencoder (SAE) approach for classification of clinical metabolomic data. This new method has the advantage of providing a confidence score for each prediction thanks to a softmax classifier and a meaningful latent space visualization and to include a new efficient feature selection method, with a structured constraint, which allows for biologically interpretable results.

Results: Experimental results on three metabolomics datasets of clinical samples illustrate the effectiveness of our SAE and its confidence score. The supervised autoencoder provides an accurate localization of the patients in the latent space, and an efficient confidence score. Experiments show that the SAE outperforms classical methods (PLS-DA, Random Forests, SVM, and neural networks (NN)). Furthermore, the metabolites selected by the SAE were found to be biologically relevant.

Conclusion: In this paper, we describe a new efficient SAE method to support diagnostic or prognostic evaluation based on metabolomics analyses.

1 Background

Deep neural networks have proven their effectiveness in bioinformatics for classification and feature selection [1, 2, 3, 4, 5]. They have also been recently used in metabolomic studies [6, 7, 8, 9, 10]. Classical stacked autoencoders [11] were used recently in metabolomic studies [12].

Autoencoders were introduced within the field of neural networks decades ago, their most efficient application at the time being dimensionality reduction [13, 14]. Autoencoders have also been used for denoising different types of data [11] to extract relevant features. One of the main advantages of the autoencoder is the projection of the data in the low dimensional latent space.

These autoencoder models include variational autoencoders (VAE) [15]. VAE networks encourage the latent space to fit a prior distribution, like a Gaussian. This can alter the accuracy of the model. In order to cope with this issue, some recent papers have proposed latent spaces with more complex distributions (e.g. mixtures of Gaussians [16]) on the latent vectors, but they are non-adaptive and unfortunately may not match the specific data distribution.
In this work, we relaxed the parametric distribution assumption on the latent space to learn a non-parametric data distribution of clusters [17]. Our network encourages the latent space to fit a distribution learned with the clustering labels rather than a parametric prior.

Recent untargeted metabolomic methods using liquid chromatography coupled with high resolution mass spectrometry (LC-MS/MS) allow for fast and high-resolution detection of massive amounts of metabolites. Metabolomics is a very promising omics field for fundamental research in biology as well as for clinical research applications. Indeed, metabolomics can be used to reveal new biomarkers of physiological or pathological states [18, 19, 20, 21], and could be particularly useful for personalized medicine [22, 23].
In this study, we described a new SAE method using structured constraints and compare its performances to classical machine learning and Neural Network methods, when applied to three clinical metabolomic databases.

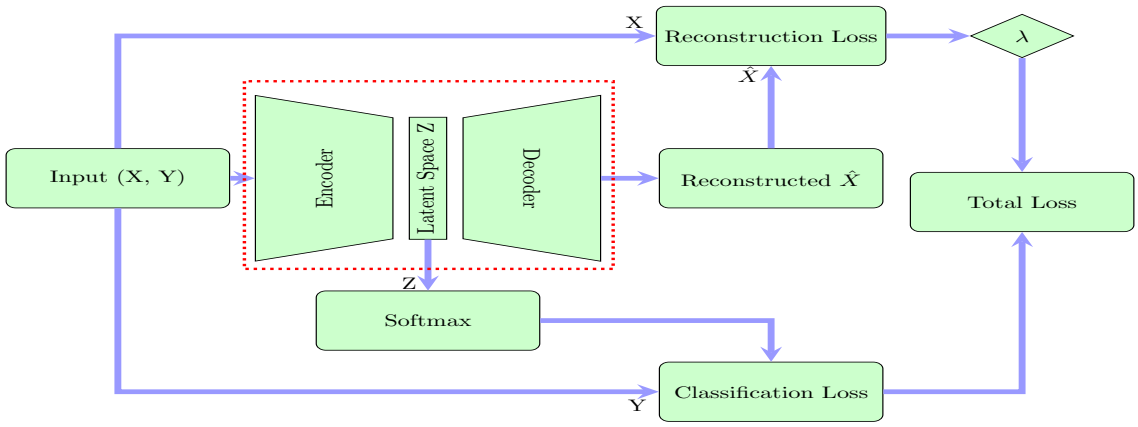

Figure 1: Supervised Autoencoder framework

2 Methods

2.1 A New supervised Autoencoder (SAE) framework

Projecting the samples in the lower dimension latent space is crucial to separate them accurately. Herein we propose to use a neural network autoencoder framework.

Let us recall that the encoder part of the autoencoder maps feature-points from a high dimensional space to a low dimensional latent space, and that the decoder maps feature points from that low dimensional space to a high dimensional space.

Figure 1 depicts the main constitutive blocks of our proposed approach. We have added to our SAE a "soft max" block to compute the classification loss.

Let X be the dataset, as an $m \times d$ data matrix made of m line samples $x_1, \ldots, x_m$. Let $y_i = j, j \in [1...k]$ be the label, indicating that the sample x_i belongs to the j-th cluster. Let Z, be the latent space, $\widehat{X}$ the reconstructed data (Figure 1) and W the weights of the neural network.

The goal is to compute the weights W minimizing the total loss, which depends on both the classification loss and the reconstruction loss. Thus, we propose to minimize the following criterion to compute the weights W of the autoencoder (see [17] for details).

$$Loss(W) = \phi(Z, Y) + \lambda\psi(\widehat{X} - X) \text{ s.t. } \|W\|_1^1 \leq \eta. \tag{1}$$

Where $\phi(Z, Y)$ is the classification loss in the latent space and $\psi(\widehat{X} - X)$ is the reconstruction loss.

The parameter λ controls the weight of the reconstruction loss in the criterion. We used the Cross Entropy Loss for the classification loss function ϕ. We used the robust Smooth ℓ_1 (Huber) Loss [24] as the reconstruction loss function ψ, as it is more robust to outliers than the classical Mean Squared Error (MSE) loss. The dimension of the latent space is defined by the number of clusters.

2.2 Structured constraints, sparsity and feature selection

The basic idea for feature selection is to use a sparse regularizer that forces some coefficients to be zero. To achieve feature selection, classically, the Least Absolute Shrinkage and Selection Operator (LASSO) formulation [25, 26, 27, 28, 29] is used to add an ℓ_1 penalty term to the classification loss. However the LASSO is computationally expensive [26, 27]. Thus, we used a feature selection method by optimizing a criterion under constraints [30].

Let us recall that the classical ℓ_2 norm constraint does not induce any sparsity. Moreover the "group Lasso $\ell_{2,1}$ constraint" induces small sparsity [31] and the ℓ_1 constraint induces unstructured sparsity [32, 33]. Thus we used $\ell_{1,1}$ constrained regularization penalty $\|W\|_1^1 \leq \eta$ for feature selection [17].

2.2.1 Algorithm

We compute the $\ell_{1,1}$ constraint with the following algorithm: we first compute the radius t_i and then project the rows using the ℓ_1 adaptive constraint t_i.

Following the work developed by [34], which proposed a double descent algorithm, we replaced the thresholding by our $\ell_{1,1}$ projection and devised a new double descent algorithm (See Barlaud and Guyard 2020 [35]) as follows :

2.3 Implementation

2.3.1 Pytorch implementation of our supervised autoencoder

We implemented our sparse supervised autoencoder model in the Pytorch framework. The losses are averaged across observations for each batch. We chose the ADAM optimizer [36], as the standard optimizer in PyTorch. We used the Netbio SAE, a linear fully connected network (LFC), which has an input layer of d neurons, 1 hidden layer of 96 neurons followed by a ReLU activation function, and a latent layer of dimension 2 (the number of classes). The parameter η is determined by the maximum accuracy after cross-validation.

We compared the Netbio SAE with a classical linear fully connected Neural Network (NN) with the same structure.

Algorithm 1 Projection on the $\ell_{1,1}$ norm—$proj_{\ell_1}(V, \eta)$ is the projection on the ℓ_1-ball of radius η, $\nabla\phi(W, M_0)$ is the masked gradient with binary mask M_0, f is the ADAM optimizer, γ is the learning rate

Input: W, γ, η
for $n = 1, \ldots, N(epochs)$ **do**
  $V \leftarrow f(W, \gamma, \nabla\phi(W))$
end for
$t := proj_{\ell_1}((\|v_i\|_1)_{i=1}^d, \eta)$
for $i = 1, \ldots, d$ **do**
  $w_i := proj_{\ell_1}(v_i, t_i)$
end for
Output: W, M_0
Input: W
for $n = 1, \ldots, N(epoch)$ **do**
  $W \leftarrow f(W, \gamma, \nabla\phi(W, M_0))$
end for
Output: W

We used the captum package [37] to compute the feature weights of the SAE.

We provide comparisons with a PLS-DA using 4 components, with Random Forests using 400 estimators and a maximum depth of 3 (using the Gini importance (GI) for feature ranking), and with a support vector classifier (SVM) with a linear kernel. For the SVM, we perform a cross-validation grid search to find the best regularization parameter C.
We provide the statistical evaluation (Accuracy, AUC, and F1 score) using a 4-fold cross validation process: the dataset is randomly divided into four parts, and trained on three of the four splits. The metrics are computed on the remaining test split, which wasn't used during training. We then repeat this process three more times, leaving a different split as the test set each time. The final metrics given in this paper are averages over the four cross-validation steps, over three different random seeds (12 different testing/training splits in total).

We compare the performances of the different methods using the F1 Score. The F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. The F1 score is more relevant than accuracy, especially for unbalanced datasets.
The computation of the statistical metrics, the classifiers, the cross-validation function and the grid search were all provided by the scikit-learn machine learning python package. The python code is available on github: `https://github.com/CyprienGille/Supervised-Autoencoder`.

2.3.2 Diagnosis with confidence score

One of the main advantages of an autoencoder is the projection of the data in the latent space, which can easily be visualized if the latent space is of dimension 2. [1] Thanks to this, we propose a clinical diagnosis simulation: having trained a network on a database of patients, we can predict a diagnosis with a confidence score for new patients. To perform this simulation, we removed a patient from each of the k classes from the databases. We then trained the SAE on (n-k) patients and we fed the k "test" patients through the best net. We thus obtained a visualization of the projections of these new "test" patients in the latent space as well as their classification with a confidence score (see figures 4, 10 and 7).

The clinician then has an accurate and reliable system to help with the diagnosis. Indeed, in addition to obtaining the confidence score for the diagnosis, the clinician can see where the patient is located among the others in the database and have a critical evaluation of the prediction (the clinician can easily see if a patient stands out).

2.4 Evaluation on 3 clinical metabolomics databases

The SAE was tested on three different metabolomic datasets : the "LUNG" , "BREAST", and "BRAIN" datasets.

The LUNG dataset was published by Mathe et al [38] and is available at MetaboLights (study identifer MTBLS28). It includes metabolomics data concerning urine samples from 469 Non-Small Cell Lung Cancer (NSCLC) patients prior to treatment and 536 controls collected from 1998 to 2007 in seven hospitals and in the Department of Motor Vehicles (DMV) from the greater Baltimore, Maryland area. Urine samples were analyzed using an unbiased metabolomics LC-MS/MS approach. Mathe et al. used Random Forests to classify patients as lung cancer patients or controls [38]. The aim was to create a new screening test for lung cancer, based on metabolomics data from urine. Lung cancer is one of the most common cancers and it is well established that early diagnosis is crucial for treatment. An efficient screening method based on urinary metabolomics could be of great benefit.

The BREAST dataset was kindly provided by Dr. Jan Budczies and can be found in the supplementary material of Budczies et al [39]. It includes metabolomics data concerning 271 breast tumor samples: 204 tumors with over-expression of estrogen receptors (ER) and 67 tumors without over-expression of ER. Metabolomics analysis was performed using Gas chromatography followed by time of flight mass spectrometry as described in [40].

[1] If the latent space is of dimension $k > 2$, we can project the latent space on a 2D plot using a PCA.

The BRAIN dataset was obtained through a study performed in our lab[1*]. It includes metabolomic data obtained on 88 frozen samples of glial tumors. The samples were retrospectively collected from two declared biobanks from the Central Pathology Laboratory of the Hospital of Nice and from the Center of Biological Resources of Montpellier (Plateforme CRB-CHUM). Consent or non-opposition was verified for every participant. Tumors were analyzed using Liquid Chromatography coupled to tandem Mass Spectrometry (LC-MS/MS) in an unbiased metabolomics approach. The details of the analysis are available in Additional file 1.

With this dataset, the goal was to create a model that accurately discriminated between mutated isocitrate dehydrogenase (IDH) and IDH wild-type glial tumors. The dataset includes (38 IDH wild-type tumors and 50 IDH-mutant tumors). This mutation is a key component of the World Health Organization classifcation of glial tumors [29]. The mutational status is usually assessed by IDH1 (R132H)-specifc (H09) immunohistochemistry. Yet this technique can lead to False-Negative results, which can only be identified by sequencing. Thus an accurate metabolomic based test, able to assess the IDH mutational status, could be a promising additional diagnostic tool.

The characteristics of the three metabolomic datasets are presented in Table 1. We chose to study these databases for their diversity both in terms of the number of features and number of patients, to test the robustness of our method on different types of databases.
The LUNG dataset includes a very large number of patients (1,005), with an equivalently large number of features (2,944), and 2 classes. The BREAST dataset includes a midsize number of patients (271), with a small number of features (161), and 2 classes. The BRAIN dataset includes a limited number of patients (88), with a much higher number of features (7,022), and 2 classes.

Table 1: Overview of the datasets.

Dataset	No. of Samples	No. of features	Sample type
LUNG	1,005	2,944	Urine
BREAST	271	161	Tumor tissue
BRAIN	88	7,022	Glial tumor tissue

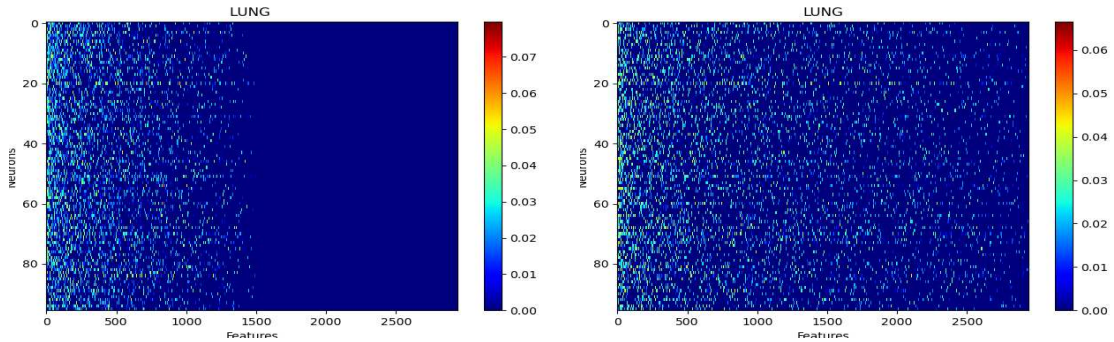

Figure 2: **LUNG** SAE Netbio Matrix: features versus hidden layer:Left with $\ell_{1,1}$ constraint,Right with ℓ_1 constraint

3 Results

3.1 LUNG dataset

3.1.1 Statistical performances

As shown in Table 2 our SAE outperformed PLS-DA, Random Forests, SVM and NN by 4.58, 9.58, 9.63 and 2.74% respectively for the F1 score. Note that we checked that increasing the number of trees for Random forests from 100 to 400 resulted in a small improvement in accuracy of only 1% while the computational cost increased by a factor of 3. The performances of the SAE were a little better when using an ℓ_1 loss than when using an ℓ_2 loss.

Table 2: **LUNG** dataset: Accuracy using 3 seeds and 4-fold cross validation: comparison with PLS-DA, Random Forest, SVM and NN

Lung	SAE ℓ_1	SAE ℓ_2	PLS-DA	RF	SVM	NN
Accuracy $\%$	81.22	80.46	76.56	72.47	76.26	78.27
AUC	80.98	80.29	76.85	74.46	78.37	77.94
F1 score	80.74	80.29	76.16	71.16	71.11	78.00

3.1.2 Feature selection using the $\ell_{1,1}$ structured constraint

Figure 2 shows the matrix ($d \times n$) of the network connections between the input layer (d feature neurons) and the hidden layer (n neurons).

It shows the benefit of using the $\ell_{1,1}$ constraint: The $\ell_{1,1}$ constraint selects features while the constraint ℓ_1 selects only weights of features. All the following results are given with the $\ell_{1,1}$ constraint.

As shown in Table 3, all methods selected metabolite "MZ 264.121", which most likely corresponds to creatine riboside (expected m/z value in the positive mode: 264.1190). Note that the SVM selected metabolite "MZ 264.121" at rank 3. Metabolite "MZ 308.098", which most likely corresponds to N-acetylneuraminic acid, was only selected by the SAE and the NN at rank 2 and 3, respectively. These metabolites were described by Mathé et al. [38] as the most important metabolites to discriminate between lung cancer patients and healthy individuals. Note that the author of RF proposes

Table 3: Top 5 features on the **LUNG** dataset. From left to right: SAE, PLS-DA, Random Forest, SVM and NN

SAE	PLS-DA	Random Forest	SVM	NN
MZ 264.12	**MZ 264.12**	**MZ 264.12**	MZ 170.06	**MZ 264.12**
MZ 308.09	MZ 126.90	MZ 441.16	MZ 126.90	MZ 126.90
MZ 126.90	MZ 613.35	MZ 584.26	**MZ 264.12**	**MZ 308.09**
MZ 232.03	MZ 170.06	MZ 486.25	MZ 94.06	MZ 613.35
MZ 332.09	MZ 243.10	MZ 204.13	MZ 110.99	MZ 332.09

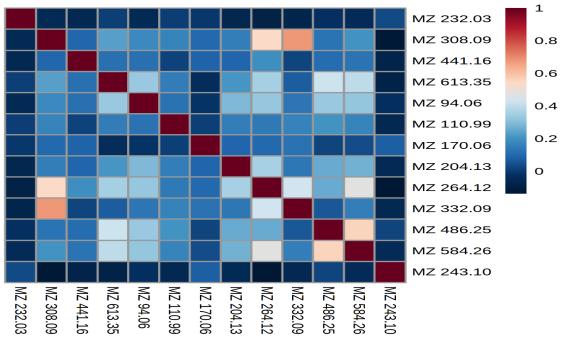

Figure 3: Correlation matrix of selected features in the **LUNG dataset**

two measures for feature ranking, the variable importance (VI) and Gini importance (GI): a recent study showed that if predictors are categorical, or real with multimodal Gaussian distributions, both measures are biased [41].

As shown in Figure 3, selected features were not significantly correlated. The highest correlation found was between MZ 308.09 and MZ 332.09, with a Pearson coefficient of 0.67. Both features correspond to adducts of N-acetylneuraminic acid (MZ 308.09 being the [M+H]+ adduct and MZ 332.09 the [M+Na]+ adduct).

3.1.3 Diagnosis in the latent space with a confidence score

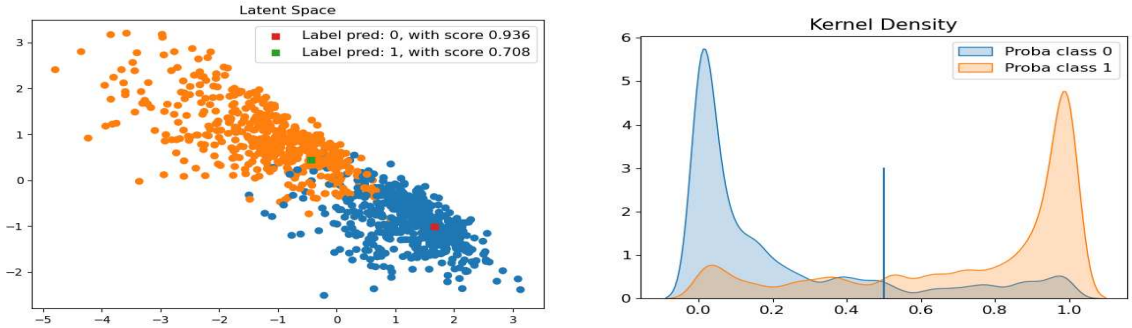

Figure 4: **LUNG** dataset. Right: Latent space, with test patients as squares. Left: Distribution using a Gaussian kernel

As shown in Figure 4, the two classes are well separated in the latent space of the SAE. Furthermore, the red and green squares show the location of the two random ”test” patients in the SAE’s latent space. The red patient is at the heart

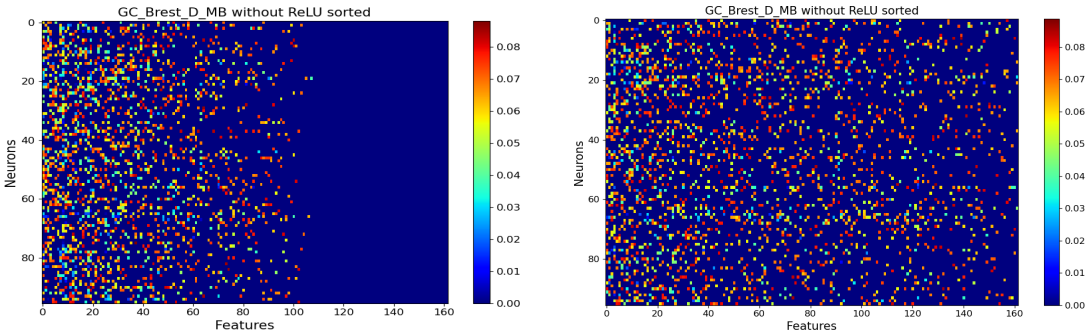

Figure 5: **BREAST** SAE Netbio Matrix: features versus hidden layer:Left with $\ell_{1,1}$ constraint,Right with ℓ_1 constraint

of the class distribution and the green patient is close to the edge of the other class. This is important for a clinician's assessment of the result. Moreover, the distribution plot shows the nearly perfect separability of the distributions calculated with the SAE, which means most of the patients were diagnosed with a high degree of confidence. The patient represented by the red square was classified in class 0 with a confidence score of 0.94 and the patient represented by the green square was labeled class 1 with a confidence score of 0.70. Both predicted labels were correct.

3.2 BREAST dataset

3.2.1 Statistical performances

Table 4: **BREAST** dataset: Accuracy using 3 seeds and 4-fold cross validation: comparison with PLS-DA, Random Forest, Logistic Regression, SVM and NN

Breast	SAE ℓ_1	SAE ℓ_2	PLS-DA	RF	SVM	NN
Accuracy %	90.15	89.05	86.58	80.23	83.20	89.04
AUC %	84.88	81.62	83.07	88.02	77.64	80.34
F1 Score	85.17	83.66	76.01	71.07	76.06	82.94

As shown in Table 4 our SAE outperformed PLS-DA, Random Forests, SVM and NN by 9.16, 14.1, 9.11 and 2.23% respectively for the F1 score. The performances of the SAE were a little better when using an ℓ_1 loss than when using an ℓ_2 loss.

3.2.2 Feature selection using the $\ell_{1,1}$ structured constraint

Figure 5 shows the matrix ($d \times n$) of the network connections between the input layer (d feature-neurons) and the hidden layer (n neurons). It shows the benefit of using the $\ell_{1,1}$ constraint: The $\ell_{1,1}$ constraint selects features, while the constraint ℓ_1 selects only weights of features.

As shown in Table 5, the SAE and the NN selected the same top five metabolites (beta-alanine, xanthine, uracil, glutamic

Table 5: Top 5 features on the **BREAST** dataset. From left to right: SAE, PLS-DA, Random Forest, SVM and NN

SAE	PLS-DA	Random Forest	SVM	NN
beta-alanine	beta-alanine	beta-alanine	3-phosphoglycerate	beta-alanine
xanthine	xanthine	xanthine	beta-alanine	xanthine
uracil	nicotinamide	glutamic acid	uracil	2-hydroxyglutaric
glutamic acid	isothreonic acid	idonic acid NIST	taurine	uracil
2-hydroxyglutaric acid	creatinine	uracil	2-ketoadipic acid	glutamic acid

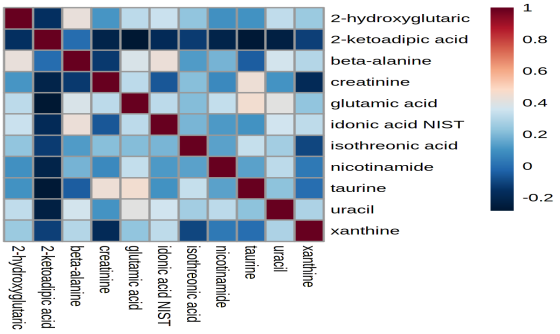

Figure 6: Correlation matrix of selected features in the **BREAST dataset**

acid). These metabolites have already been shown to have significantly different concentrations in ER breast tumors compared to ER+ breast tumors in the original paper by Budczies et al [39]. Increased concentrations of glutamic acid and 2-hydroxyglutaric acid indicate higher glutaminolysis, a key feature of metabolic changes in cancer cells. As shown in Budczies et al [39], increased concentrations of uracil, xanthine and beta-alanine levels are related to higher hexokinase 3, xanthine dehydrogenase and 4-aminobutyrate aminotransferase expressions, respectively.

As shown in Figure 6, selected features were highly correlated.

3.2.3 Prognosis in the latent space with confidence score

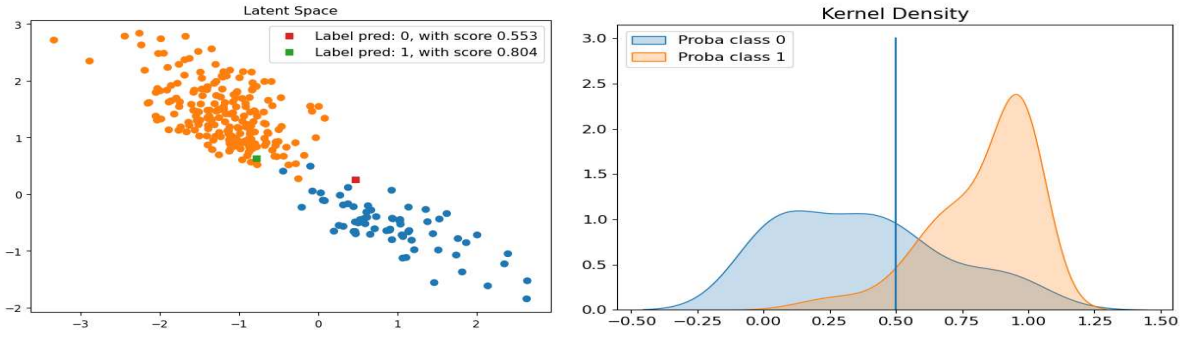

Figure 7: **BREAST** dataset. Left: latent space of the SAE. Right: Distribution using a Gaussian Kernel

Figure 7 (left), shows the accurate separation of the two classes in the latent space of the SAE. The red and green squares show the location of the two random ”test” patients in the SAE’s latent space. The patient represented by the red square

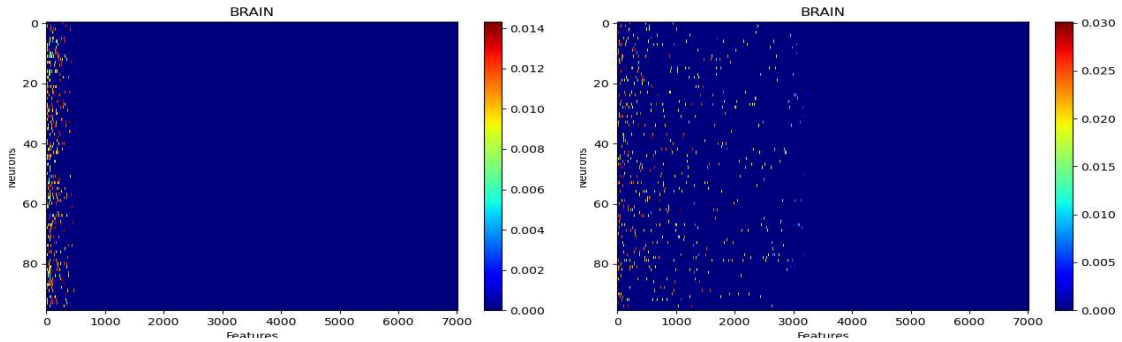

Figure 8: **BRAIN** SAE Netbio Matrix: features versus hidden layer: Left with $\ell_{1,1}$ constraint, Right with ℓ_1 constraint

was classified in class 0 with a confidence score of 0.55 and the patient represented by the green square was labeled class 1 with a confidence score of 0.80. Both predictions are correct. Figure 7 (right) shows the separability of the distributions calculated with the SAE.

3.3 BRAIN dataset

3.3.1 Statistical performances

Table 6: **BRAIN** dataset Accuracy using 3 seeds and 4-fold cross validation: comparison with PLS-DA, Random Forest , SVM and NN

Brain	SAE ℓ_1	SAE ℓ_2	PLS-DA	RF	SVM	NN
Accuracy $\%$	92.80	88.63	84.84	86.73	87.12	75.75
AUC $\%$	93.29	88.64	85.37	89.5	87.52	74.85
F1 score	92.66	88.40	83.88	88.05	86.51	74.19

Table 6 shows that, despite the small number of patients, the supervised autoencoder outperformed PLS-DA, Random Forest, SVM and NN by 8.78, 4.61, 6.15 and 18.47% respectively for the F1 score. For this base with few patients the performance of NNs collapses as reported in the literature. As for the other databases, the performances of the SAE were a little better when using an ℓ_1 loss than when using an ℓ_2 loss.

3.3.2 Feature selection using the $\ell_{1,1}$ structured constraint

Figure 8 shows the matrix ($d \times n$) of the network connections between the input layer (d feature-neurons) and the hidden layer (n neurons). It shows the benefit of using the $\ell_{1,1}$ constraint: The $\ell_{1,1}$ constraint selects features, while the constraint ℓ_1 selects only weights of features.

As expected, the top features selected by each method (shown in Table 7) correspond mainly to different isotopes and adducts of 2-hydroxyglutarate (marked in bold). The features selected using the SAE were all different adducts of

Table 7: **BRAIN** dataset with 7,022 features : Top 5 features selected by the SAE, PLS-DA, Random Forests, SVM and NN

SAE	PLS-DA	RF	SVM	NN
NEG MZ147.028	POS MZ131.034	NEG MZ148.031	POS MZ132.523	NEG MZ148.031
POS MZ132.037	POS MZ132.523	NEG MZ215.016	NEG MZ147.028	NEG MZ147.028
POS MZ171.026	POS MZ132.037	POS MZ132.037	POS MZ131.034	POS MZ132.037
POS MZ132.037	NEG MZ147.028	POS MZ85.029	POS MZ132.037	POS MZ85.029
POS MZ149.044	POS MZ171.026	POS MZ132.523	POS MZ171.026	POS MZ173.030

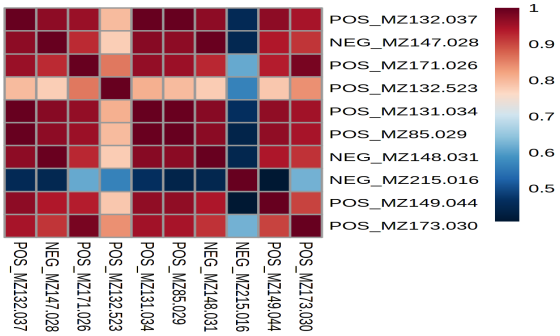

Figure 9: Correlation matrix of selected features in the **BRAIN dataset**

this specific product of IDH-mutated cells. Indeed, POS_MZ132.03 and POS_MZ131.03 correspond to the [M+H-H2O]+ adduct of 2-hydroxyglutarate with one 13C isotope for the first ion. POS_MZ171.02 is the [M+Na]+ adduct, NEG_MZ147.02 is the [M-H]- and POS_MZ86.03 is the [M+Na+H]2+ adduct. NEG_MZ148.03 is the [M-H]- adduct of 2-hydroxyglutarate with one 13C isotope. POS_MZ173.03 is the [M+Na]+ adduct with two 13C isotope. Finally, POS_MZ149.04 is the [M+H]+ adduct ion of 2-hydroxyglutarate. As expected, and shown in Figure 9, these features, all corresponding to adducts of 2-hydroxyglutarate, were highly correlated.

3.3.3 Diagnosis in the latent space with confidence score

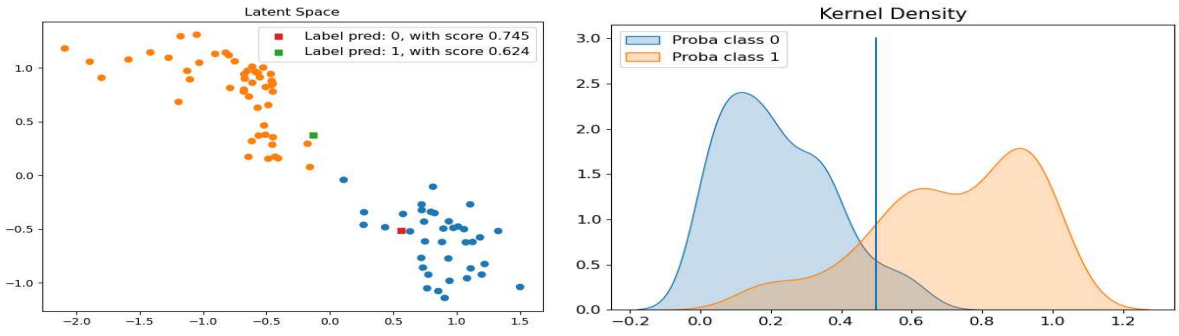

Figure 10: **BRAIN** dataset. Left: latent space of the SAE, Red and green squares are "test" patients. Right: Distribution using a Gaussian kernel

Figure 10 (left), shows the nearly perfect separation of the two classes in the latent space of the SAE. Furthermore, the red and green squares show the location of the two random ”test” patients in the SAE's latent space. The patient represented by the red square was classified in class 0 with a confidence score of 0.75 and the patient represented by the green square was labeled class 1 with a confidence score of 0.62. Both predictions were correct. Figure 10 (right) shows the peak separability of the distributions calculated with the SAE. It shows that most patients will have a good prediction with a high degree of confidence.

4 Discussion

Thus, we have shown that our SAE outperformed classical machine learning methods and NN for classification of metabolomics data, while providing reliable confidence score for the predictions and performing relevant feature selection.

The real distributions of many datasets, including metabolomics datasets, are far more complex than multi-gaussian mixtures.Thus we chose to use a non-parametric supervised autoencoder (SAE) rather than a classical autoencoder that assumes a latent space modeling [42, 43] and force a multi-gaussian distribution upon the data.

Regardless of data size and feature space dimensions, the SAE outperforms all other methods (PLS-DA, Random Forests, SVM and NN). As expected, the NN also outperformed classical methods (PLS-DA, Random Forests and SVM), except on small databases. Indeed, NN are known to be less accurate when trained on small numbers of samples [44, 45]. Furthermore, as anticipated, the SAE's performances were a little better when using the Huber loss than when using the MSE. This is most likely due to the fact that the Huber loss is more robust to outliers.

The SAE provides high-level distribution visualization of the samples in the latent space, as well as their classification confidence score. This is crucial for any diagnostic tool. Indeed, these two features enable clinicians to gauge how reliable each prediction is and if a sample corresponds to a potential outlier, for which predictions should be considered with particular care.

Metabolomics is a very promising approach, particularly adapted to routine clinical practice, because metabolomics analyses are fast and relatively inexpensive. However, human metabolomics are complex data, influenced by many external and internal factors. The high number of features included in metabolomics analyses require high performance statistical methods such as our SAE to be exploited. However, no statistical method can replace the critical reasoning of a researcher to make conclusions on the statistical results and to identify potential confounding factors. To make such

conclusions, the statistical method needs to have some degree of interpretability.

Interestingly, the SAE combined with a structured projection provides efficient feature selection (Tables 3, 5 and 7). This feature selection step is crucial for interpretability. Better yet, we have verified that the selected features in the LUNG, BREAST and BRAIN datasets were known to be biologically relevant metabolites. Efficient feature selection adds interpretability to the model which is crucial for metabolomic studies in biological research or clinical trials.

We have observed that selected features can have a low to very high degree of correlation. In our case, the correlated features were isotopes and adducts of metabolites with high weights for the classification. Even though multivariate methods, such as the one we have used, account for correlation, correlated features do have an impact on feature selection and the performances of the trained models. When studying metabolomics one must adapt the level of filtering. Indeed, filtering removes isotopes and adducts but can also remove important features. This must be taken into consideration when using our SAE or any other classification method for metabolomics analyses.

5 Conclusion

In this paper we have proposed a new and efficient classification method for metabolomics datasets, based on the representation of data on the latent space of a new supervised autoencoder (SAE). In clinical applications, our method provides a diagnosis score for each patient's predicted class. Moreover, from a statistical point of view (Accuracy, AUC, F1 score) our SAE outperformed PLS-DA, Random Forest, SVM, and NN while selecting biologically relevant features.

References

[1] Song, Y., Zheng, S., Li, L., Zhang, X., Zhang, X., Huang, Z., Chen, J., Wang, R., Zhao, H., Zha, Y., Shen, J., Chong, Y., Yang, Y.: Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2021)

[2] Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.-Z.: Deep learning for health informatics. IEEE Journal of Biomedical and Health Informatics **21**(1), 4–21 (2017)

[3] Min, S., Lee, B., Yoon, S.: Deep learning in bioinformatics. Briefings in Bioinformatics **18**(5), 851–869 (2016)

[4] Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., Tao, Y., Guo, Y., Ni, X., Shi, T.: Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Frontiers in Genetics **9**, 477 (2018)

[5] Sen, P., Lamichhane, S., Mathema, V.B., McGlinchey, A., Dickens, A.M., Khoomrung, S., Orešič, M.: Deep learning meets metabolomics: a methodological perspective. Briefings in Bioinformatics (2020)

[6] Alakwaa, F., Chaudhary, K., Garmire, L.: Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data. Journal of Proteome Research, **17**, 337–347 (2018)

[7] Bradley, W., Robert, P.: Multivariate analysis in metabolomics. Current Metabolomics **1**, 92–107 (2013)

[8] Asakura, P., Date, Y., Kikuchi, J.: Application of ensemble deep neural network to metabolomics studies. Analytica Chimica Acta **1037**, 92–107 (2018)

[9] Mendez, K., Broadhurst, D., Reinke, S.: Application of artificial neural networks in metabolomics: A historical perspective. Metabolomics **15** (2019)

[10] Sen, P., Lamichhane, S., Mathema, V.B., McGlinchey, A., Dickens, A.M., Khoomrung, S., Orešič, M.: Deep learning meets metabolomics: a methodological perspective. Briefings in Bioinformatics **22**(2), 1531–1542 (2020)

[11] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010)

[12] Xiaojing, F., Xiye, W., Mingyang, J., Zhili, P., Shicheng, Q.: An improved stacked autoencoder for metabolomic data classification. Hindawi Computational Intelligence and Neuroscience **2021** (2021)

[13] Hinton, Z.R. Geoffrey: Autoencoders, minimum description length and helmholtz free energy. In: Advances in Neural Information Processing Systems, pp. 3–10 (1994)

[14] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning vol. 1. MIT press, ??? (2016)

[15] Kingma, D., Welling, M.: Auto-encoding variational bayes. International Conference on Learning Representation (2014)

[16] Dilokthanakul, N., Mediano, P.A.M., Garnelo, M., Lee, M.C.H., Salimbeni, H., Arulkumaran, K., Shanahan, M.: Deep unsupervised clustering with gaussian mixture variational autoencoders (2016). 1611.02648

[17] Barlaud, M., Guyard, F.: Learning a sparse generative non-parametric supervised autoencoder. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, TORONTO , Canada (2021)

[18] Yazdani, H., Cheng, L.L., Christiani, D.C., Yazdani, A.: Bounded fuzzy possibilistic method reveals information about lung cancer through analysis of metabolomics. IEEE/ACM Transactions on Computational Biology and Bioinformatics **17**(2), 526–535 (2020)

[19] Liu, Y., Xu, X., Deng, L., Cheng, K.-K., Xu, J., Raftery, D., Dong, J.: A novel network modelling for metabolite set analysis: A case study on crc metabolomics. IEEE Access **8**, 106425–106436 (2020)

[20] Banimustafa, A., Hardy, N.: A scientific knowledge discovery and data mining process model for metabolomics. IEEE Access **8**, 209964–210005 (2020)

[21] Qi, Z., Voit, E.O.: Strategies for comparing metabolic profiles: Implications for the inference of biochemical mechanisms from metabolomics data. IEEE/ACM Transactions on Computational Biology and Bioinformatics **14**(6), 1434–1445 (2017)

[22] Long, N.P., Nghi, T.D., Kang, Y.P., Anh, N.H., Kim, H.M., Park, S.K., Kwon, S.W.: Toward a Standardized Strategy of Clinical Metabolomics for the Advancement of Precision Medicine. Metabolites **10**(2), 51 (2020). doi:10.3390/metabo10020051. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. Accessed 2020-12-01

[23] Cakmak, A., Celik, M.H.: Personalized metabolic analysis of diseases. IEEE/ACM Transactions on Computational Biology and Bioinformatics **18**(3), 1014–1025 (2021)

[24] Huber, P.J.: Robust statistics. 1981

[25] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288 (1996)

[26] Hastie, T., Rosset, S., Tibshirani, R., Zhu, J.: The entire regularization path for the support vector machine. Journal of Machine Learning Research **5**, 1391–1415 (2004)

[27] Friedman, J., Hastie, T., Tibshirani, R.: Regularization path for generalized linear models via coordinate descent. Journal of Statistical Software **33**, 1–122 (2010)

[28] Hastie, T., Tibshirani, R., Wainwright, M.: Statistcal learning with sparsity: The lasso and generalizations. CRC Press (2015)

[29] Li, J., Cheng, K., Wang, S., Morstatter, F., P. Trevino, R., Tang, J., Liu, H.: Feature selection: A data perspective. ACM Computing Surveys **50** (2016). doi:10.1145/3136625

[30] Barlaud, M., Belhajali, W., Combettes, P., Fillatre, L.: Classification and regression using an outer approximation projection-gradient method, vol. 65, pp. 4635–4643 (2017)

[31] Barlaud, M., Chambolle, A., Caillau, J.-B.: Classification and feature selection using a primal-dual method and projection on structured constraints. International Conference on Pattern Recognition, Milan (2020)

[32] Condat, L.: Fast projection onto the simplex and the l1 ball. Mathematical Programming Series A **158**(1), 575–585 (2016)

[33] Perez, G., Barlaud, M., Fillatre, L., Régin, J.-C.: A filtered bucket-clustering method for projection onto the simplex and the ℓ_1-ball. Mathematical Programming (2019)

[34] Zhou, H., Lan, J., Liu, R., Yosinski, J.: Deconstructing lottery tickets: Zeros, signs, and the supermask. In: Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 3597–3607. Curran Associates, Inc., ??? (2019)

[35] Barlaud, M., Guyard, F.: Learning sparse deep neural networks using efficient structured projections on convex constraints for green ai. International Conference on Pattern Recognition, Milan (2020)

[36] Kingma, D., Ba, J.: a method for stochastic optimization. International Conference on Learning Representations, 1–13 (2015)

[37] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Neural Information Processing Systems, Barcelone, Spain **30** (2017)

[38] Mathé *et al*, E.: Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. Cancer research **74**(12), 3259–3270 (2014)

[39] Budczies, J., Brockmöller, S., Müller, B., Barupal, D., Richter-Ehrenstein, C., Kleine-Tebbe, A., Griffin, J., Orešič, M., Dietel, M., Denkert, C., Fiehn, O.: Comparative metabolomics of estrogen receptor positive and estrogen receptor negative breast cancer: Alterations in glutamine and beta-alanine metabolism. Journal of Proteomics **94**, 279–288 (2013)

[40] Budczies, J., Denkert, C., Müller, B.M., Brockmöller, S.F., Klauschen, F., Györffy, B., Dietel, M., Richter-Ehrenstein, C., Marten, U., Salek, R.M., Griffin, J.L., Hilvo, M., Orešič, M., Wohlgemuth, G., Fiehn, O.: Remodeling of central metabolism in invasive breast cancer compared to normal breast tissue – a GC-TOFMS based metabolomics study. BMC Genomics **13**(1), 334 (2012). doi:10.1186/1471-2164-13-334. Accessed 2022-06-24

[41] Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340–1347 (2010)

[42] Emdadi, A., Eslahchi, C.: Auto-HMM-LMF: feature selection based method for prediction of drug response via autoencoder and hidden Markov model. BMC Bioinformatics (2021)

[43] Liu, D., Huang, Y., Nie, W., Zhang, J., Deng, L.: SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost. BMC Bioinformatics **22** (2021)

[44] Markham, I.S., Rakes, T.R.: The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression. Computers & Operations Research **25**(4), 251–263 (1998). doi:10.1016/S0305-0548(97)00074-9. Accessed 2022-06-24

[45] Hush: Classification with neural networks: a performance analysis, 277–280 (1989). doi:10.1109/ICSYSE.1989.48672

V. EXEMPLE DE DEVELOPPEMENT D'UN TEST DIAGNOSTIQUE BASE SUR LA METABOLOMIQUE

La classification des tumeurs gliales de l'adulte évolue régulièrement [58,59]. Celle-ci est basée sur des caractéristiques histologiques mais également sur des caractéristiques moléculaires telles que la présence ou l'absence d'une mutation de l'Isocitrate DésHydrogénase ou IDH. En pratique, il est parfois difficile de trancher sur le grade histologique, classé de 1 à 3 et les tests de détection de la mutation IDH sont parfois mis en défaut.
Dans ce contexte, une aide à la classification, basée sur la métabolomique, pourrait être pertinente, d'autant plus si celle si pouvait être utilisée pour des échantillons fixés en paraffine.

Nous avons utilisé une méthode de sélection de variable innovante, le BOLASSO, initialement proposée par Bach en 2008 [60] et appliquée à des données de métabolomique par Bujak et al en 2016 [30]. Le principe du BOLASSO est de réaliser un grand nombre de régression LASSO aux variables par une méthode de bootstrap. A chaque régression LASSO, les coefficients de certaines variables sont portés à zéro et seules quelques variables gardent des coefficients non nuls : ces variables sont ainsi sélectionnées. La probabilité de sélection de chaque variable peut être estimée en comptant la fréquence à laquelle son coefficient n'est pas nul. En appliquant une sélection de variable basée sur cette probabilité de sélection et non sur une unique régression LASSO, les variables sélectionnées sont plus robustes.

Nous avons utilisé le BOLASSO pour construire deux modèles de régression logistiques simples permettant de prédire le grade histologique ainsi que la présence ou l'absence de mutation IDH à partir de données de métabolomique issues d'échantillons de tumeurs gliales congelées. Nous avons ensuite montré que ces modèles pouvaient être utilisés pour prédire le grade histologique et la présence ou l'absence de mutation IDH à partir de données de métabolomique issues d'échantillons de tumeurs gliales fixées en paraffine, avec des performances diagnostiques intéressantes.

L'analyse des métabolites sélectionnés pour former ces modèles a montré qu'un unique métabolite, le 2-hydroxyglutarate (2HG), avait été sélectionné pour prédire le statut IDH et que 2 métabolites, l'acide amino-adipique (AAA) et l'acide guanidino-acetique (GAA), avait été sélectionnés pour prédire le grade histologique.

Comme mentionné précédemment, le faible nombre de métabolites sélectionnés pour créer nos modèles est lié à la méthode BOLASSO que nous avons utilisée. Cette approche semble particulièrement pertinente pour l'élaboration de tests diagnostiques à partir de données de métabolomique car elle permet de générer des modèles simples et donc plus facilement compréhensibles et reproductibles. En effet il est plus facile de se concentrer sur le rationnel biologique liant le niveau de quelques métabolites à un état pathologique plutôt que de trouver le lien entre plusieurs dizaines de métabolites. De plus, une fois ces quelques métabolites identifiés, on peut facilement les rechercher par une analyse ciblée sur de nouveaux échantillons et la simplicité du modèle limite le risque d'overfitting. En revanche, le fait que cette méthode entraine la sélection d'un nombre aussi faible de variable peut être limitant pour une utilisation plus exploratoire, notamment en biologie.

Ainsi, nous avons pu élaborer deux modèles fiables, simples et reproductibles, permettant de prédire le grade histologique et le statut mutationnel IDH à partir d'une analyse métabolomique d'échantillons de tumeurs gliales congelés ou fixés en paraffine. De plus, ces modèles sont basés sur quelques métabolites pertinents, ayant un rationnel biologique établi pour le 2HG et plausible pour l'AAA et le GAA.

Article 3 : Identification of metabolomic markers in frozen and formalin-fixed, paraffin-embedded samples of diffuse gliomas in adults

David CHARDIN, Lun JING, Mélanie CHAZAL-NGO-MAI, Jean-Marie GUIGONIS, Valérie RIGAU, Catherine GOZE, Hugues DUFFAU, Thierry VIROLLE, Olivier HUMBERT, Thierry POURCHER, Fanny BUREL-VANDENBOS

Manuscrit soumis à Brain Pathology

Title: **Identification of metabolomic markers in frozen and formalin-fixed, paraffin-embedded samples of diffuse gliomas in adults**.

David CHARDIN*MD, Lun JING*PhD, Mélanie CHAZAL-NGO-MAI MD, Jean-Marie GUIGONIS PhD, Valérie RIGAU MD, PhD, Catherine GOZE MD, PhD, Hugues DUFFAU MD, PhD, Thierry VIROLLE, PhD, Olivier HUMBERT, MD, PhD, Thierry POURCHER§ PhD, Fanny BUREL-VANDENBOS § MD, PhD.

* David Chardin and Lun Jing are co-first authors
§ Thierry Pourcher and Fanny Burel-Vandenbos are co-last uthors

Abstract

Since 2016, diffuse gliomas (DG) are classified according to several histomolecular criteria, including the gliomagenesis pathway, i.e. the presence or absence of *IDH* (isocitrate dehydrogenase) gene mutation. By applying an untargeted metabolomic approach, we aimed to identify metabolomic signatures associated with the gliomagenesis pathway (*IDH*-mutant or *IDH*-wt) and tumor grade, according to the revised 2016 WHO classification on frozen samples. We also aimed to evaluate the diagnostic performances of these signatures on Formalin-Fixed and Paraffin-Embedded (FFPE) tumor samples. An untargeted metabolomic study was performed using mass spectrometry coupled with liquid chromatography on a cohort of 213 DG samples (including 82 pairs of frozen and FFPE samples, 44 distinct FFPE samples and 5 distinct frozen samples). After metabolomic analysis, 905 distinct metabolites were found in both frozen samples and FFPE samples. Logistic regression with LASSO penalization was used on frozen samples to build classification models in order to identify *IDH*-mutant vs *IDH*-wildtype DG and high-grade (IV) vs low-grade (II & III) DG samples. These models were then tested on FFPE samples. Hydroxyglutaric acid (2HG) $[M-H2O+H]^+$ adduct was found to be a metabolite of interest to predict *IDH* mutational status. When training a model based on hydroxyglutaric acid values on frozen samples and testing this model on FFPE samples, AUC was 82.6%, sensitivity was 70.6% and specificity was 80.4%. Aminoadipic acid (AAA) and guanidinoacetic acid (GAA) were found to be significantly associated with grade. When training a model based on AAA and GAA values on frozen samples and testing this model on FFPE samples, AUC was 80%, sensitivity was 75% and specificity was 74.5%. Metabolomic data can be useful in the classification of diffuse gliomas, both in frozen and FFPE samples.

Introduction

Genetic alterations are one of the paramount mechanisms of carcinogenesis with metabolic reprogramming being one of their main consequences. Metabolite changes in a tumor can be studied via metabolomic studies [1-3], meaning approaches based on determining levels of different small molecules or metabolites in biological samples (tissue, cells, serum, urine, etc.). Different metabolomic approaches exist: targeted, based on identifying preselected metabolites within a sample, and untargeted, which detects as many metabolites as possible in a sample. The most common metabolomic techniques used in cancer research are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry with liquid chromatography (LC-MS) or gas chromatography (GC-MS). Untargeted approaches are of increasing interest because they enable the assessment of thousands of metabolites with a single analysis and can therefore be used to discover novel biomarkers of cancer. Furthermore, generated data may be used for supervised classification methods, which are emerging tools that help classify tumors along with histological assessment. To be convenient and useful in medical practice, techniques must be adapted to the most common type of tumor sample conditioning and should be retrospectively applicable for each collected sample. In research, most metabolomic studies have been performed on frozen tissues or biological fluids. However, in routine practice within pathology laboratories, representative frozen samples are inconsistently available, whereas tissue samples of all patients are Formalin-Fixed and Paraffin-Embedded (FFPE) and then conserved in archives. FFPE samples represent a very large potential for exploitable tissues and more and more techniques are adapted for use on these types of specimens. Thus, adapting metabolomic techniques to FFPE samples offers a promising prospective for their application in medical practice. Surprisingly, however, only a small number of metabolomic studies have been performed on FFPE samples [4, 5].

Diffuse gliomas (DG) are the most frequent primary malignant tumors of the central nervous system (CNS) [6]. Because of their highly invasive behavior in the brain and relative radio- and chemoresistance, these tumors remain incurable despite combinations of surgery and adjuvant

therapies. Until 2016, DGs in adults were classified according to their histological subtypes (astrocytic, oligodendroglial or mixed) and their World Health Organization grade (II (low-grade) and III and IV (high grades)), with the most aggressive tumor being the glioblastoma (GBM), defined as a WHO grade IV astrocytic tumor. DGs are genetically heterogeneous tumors. The revised version of the 2016 WHO classification of CNS tumors [7] included molecular data in the diagnosis of DG, leading to an integrative histomolecular diagnosis and increasing the reproducibility of the diagnosis among pathologists. Thus, astrocytic DG in adults must be classified according to the gliomagenesis pathway, i.e. the presence or absence of mutation in *IDH* (isocitrate dehydrogenase) genes (*IDH1* or *IDH2*). All oligodendrogliomas are *IDH*-mutant DG and a codeletion of chromosomes 1p and 19q is mandatory for their diagnosis. Most grade II and grade III astrocytic DG (80%) are *IDH*-mutant gliomas and harbor an *ATRX* mutation. GBM with *IDH* mutation usually correspond to malignant progressions of low-grade *IDH*-mutant astrocytomas (> 90%) and are so-called “secondary GBM”. *IDH*-mutant GBM are rare (< 10% GBM). Most GBM are *IDH*-wildtype (*IDH*-wt) and are called “de novo GBM”. These GBM are associated with the poorest prognosis. Only a small proportion of low-grade astrocytomas are *IDH*-wt and tend to show a dismal prognosis [8]. Of note, mixed tumors (oligo-astrocytomas) have nearly disappeared in the 2016 version of the WHO classification and must be re-classified as astrocytomas or oligodendrogliomas according to their molecular signatures [9].

Mutation in a gene coding for the enzymes IDH1 or IDH2 induces an aberrant enzymatic activity, leading to the production of an oncometabolite, 2-hydroxyglutarate (2-HG) [10]. Therefore, *IDH*-mutant DGs are enriched in 2-HG as compared to *IDH*-wt DGs [10]. Most metabolomic studies in DG have been performed on fresh or frozen tissues or liquids [1-3, 11], often in small cohorts [12], and were carried out before the revised version of the WHO classification [1-3]. Therefore, most data resulting from these studies need to be updated regarding the more recent WHO classification. Only one metabolomic study has been performed on FFPE diffuse gliomas [13]. Using a targeted metabolomic approach, Sahm et al.[13] demonstrated that 2-HG was detectable in FFPE gliomas and, as expected, was significantly increased in *IDH*-mutant DG.

By applying an untargeted metabolomic approach on a large cohort of frozen and FFPE diffuse glioma samples, we aimed to identify metabolomic signatures associated with the gliomagenesis pathway (*IDH*-mutant or *IDH*-wt) and grade, according to the revised 2016 WHO classification on frozen samples, as well as evaluate the diagnostic performances of these signatures on FFPE samples.

Methods

Sample collection

Frozen and FFPE tumor samples were collected from the Pathology Departments of Nice University Hospital and from the platform CRB-CHUM of Montpellier University Hospital (Biobank BB-0033-0031). Tissue collections were declared to the French Health Ministry (as required by French legislation). Written consent and/or non-opposition was obtained for each patient. All cases were classified according to the 4th revision of the WHO classification of CNS tumors by two expert neuropathologists (FBV and VR). Of note, all the lower-grade IDH-wt astrocytomas harbored one of the following molecular features: association of chromosome 7 gain and chromosome 10 loss, *EGFR* amplification and/or a mutation of *TERT* promoter. In this study, grade was solely based on "classical" morphological criteria (cell density, nuclear atypia, mitosis count, necrosis, and microvascular proliferation).

Sample preparation

Frozen tissues (sections of 100 µm thickness) were placed in microcentrifuge tubes and ground in 1 mL of cold methanol (LC-MS grade, Merck Millipore, Molsheim, France) using pestles. FFPE tissues (sections of 100 µm thickness) were mixed with 1 mL of methanol at 70°C for 30 min and then at 0°C for 15 min, and centrifuged at 13,000 RPM for 10 min at 0°C. The supernatant and homogenized frozen tissues were incubated overnight at -20°C then centrifuged at 15,000 RPM for 15 min. The supernatant was then removed and dried using a SpeedVac concentrator (SVC100H, SAVANT, Thermo Fiosher Scientific, Villebon-sur-Yvette, France). Lyophilized samples were resuspended in 100 µl of a 50:50 acetonitrile-H2O mix (LC-MS grade, Merck, Millipore) prior to LC-MS/MS analyses.

LC-MS/MS analysis

Frozen samples and FFPE samples were analyzed separately.

Metabolomic analyses were performed using LC-MS/MS. Liquid chromatographic analysis was performed using the DIONEX Ultimate 3000 HPLC System (Thermo Fisher Scientific). 10 µl of each sample was injected into a Synergi 4 µm Hydro-RP 80 Å, 250 x 3.0 mm column (Phenomenex, Le Pecq, France). The mobile phases were composed of 0.1% formic acid (Thermo Fisher Scientific) in water (A) and 0.1% formic acid in acetonitrile (B). The gradient was set as follows with a flow rate of 0.9 mL/min: 0% phase B from 0 to 5 min, 0-95% B from 5 to 21 min, holding at 95% B to 21.5 min, 95-0% B from 21.5 to 22 min, holding at 0% B until 25 min for column equilibration. Mass spectrometry analysis was carried out on a Q Exactive Plus Orbitrap mass spectrometer (Thermo Fisher Scientific) with a heated electrospray ionization source, HESI II, operating in both positive and negative mode. High-resolution accurate-mass full-scan MS and top 5 MS2 spectra were collected in a data-dependent fashion at a resolving power of 70,000 and 35,000 at m/z 400, respectively. All samples were successively treated in the same run.

Metabolomic profiling

Data from frozen samples and FFPE samples were processed separately.

Raw data obtained from positive and negative ionization mode were analyzed separately using MZmine (Version 2.53) [14]. Mass detection was performed using the Mass detector tool (mass detector: Wavelet transform, MS level 1; Noise level 10^5, scale level: 5, Wavelet window size: 30%). Chromatograms were detected using the ADAP chromatogram builder [15] (MS level: 1, Minimum group size in number of scans: 5, Group intensity threshold : 5x10^2, minimum highest intensity : 10^5, m/z tolerance : 10 ppm). Peaks were separated using the Peak extender module (M/Z tolerance: 10 ppm, minimum height: 10^5). Retention times were normalized using the retention time calibration module (m/z tolerance: 10 ppm; retention time tolerance (relative): 10%, minimum height: 10^5).

Peaks were then aligned using the RANSAC aligner (random sample consensus) algorithm with a tolerance of 10 ppm in m/z and 1 min in retention time. Peaks were then identified using the Human metabolome database [16] (HMDB, version 3.0) with 10 ppm of mass tolerance. Missing values were filled in using the same m/z and RT range gap filler with a tolerance of 10 ppm in m/z. The results obtained with each polarity were combined and, for metabolites that were identified in both modes, only the mode for which the peak had the highest mean intensity was considered. Furthermore, only peaks that had intensities over 10^6 in at least 30 samples were kept for analysis to eliminate noisy peaks. Finally, non-attributed values in the final databases were replaced by an arbitrary small value, corresponding to the peak intensity threshold (10^5).

After statistical analysis, metabolites of interest were individually verified (MS and MS2 spectra), using compound discoverer 3.1 (Thermo) and matching between the experimental m/z and the reference MS/MS spectrums of the m/z cloud (Thermo) and Metlin databases [17] (available as supplementary data).

Statistical analyses

Statistical analyses were carried out using R version 3.6.3 with packages “glmnet”, “pROC” and “Metabonalyze”. Raw data were mean-centered, scaled and log-transformed before performing a logistic regression with L1 penalization (LASSO).

The first step consisted of selecting relevant features using frozen samples as a training set. Logistic regression with LASSO penalization was used with 4-fold cross-validation (CV) to select the optimal value of lambda, for which maximum Area Under the Receiver Operating Characteristics Curve (AUC) was obtained. Using this method, metabolites were selected which contributed the most to classification between groups. Robustness of each selected metabolite was measured using a resampling-based bootstrap procedure, as done by Bujak et al.[18] 100 resamples were performed and metabolites were ranked according to their selection probability based on the number of times

they had been selected after 100 resamples. The mean of the performances of the models were measured as a mean AUC. The mean weights of each metabolite were also recorded.

As a second step, logistic regression models were trained using only metabolites for which the selection probability was over 85%. These models were first evaluated on frozen samples using 4-fold cross validation. Then, the reproducibility of the results was evaluated by training a logistic regression model using frozen samples and testing this model on FFPE samples. The performances of the model were calculated as AUC, Accuracy, Sensitivity and Specificity, using ROC curve analysis.

Mean values of the metabolites of interest were also compared between groups using Student's t-test.

Results

The cohort consisted of 213 tumor samples: 82 pairs of frozen and FFPE tumor samples, along with distinct 44 FFPE samples and 5 distinct frozen samples. The classification of tumor samples is detailed in table 1. Because most previous metabolomics studies were performed on fresh or frozen tissues, the results obtained in frozen samples in this study were used as a gold standard to compare the results obtained in FFPE samples.

Table 1. Histomolecular classification of the diffuse glioma samples in the study according to the WHO 2016 classification.

Histological subtypes	IDH status	WHO grade	Frozen samples n=87 (%)	FFPE samples n=126 (%)
Oligodendroglioma	Mutant	II	7 (8.0%)	15 (11.9%)
Astrocytoma	Mutant	II	19 (21.8%)	29 (23.0%)
Astrocytoma	Mutant	III	10 (11.5%)	15 (11.9%)
Glioblastoma	Mutant	IV	12 (13.8%)	16 (12.7%)
Astrocytoma	Wild type	II	9 (10.3%)	15 (11.9%)
Astrocytoma	Wild type	III	12 (13.8%)	20 (15.9%)
Glioblastoma	Wild type	IV	18 (20.7%)	16 (12.7%)

All samples were analyzed by LC-MS/MS. 3267 individual peaks were analyzed in the frozen samples, among which 919 could correspond to metabolites from the HMDB database. 2616 individual peaks

were analyzed in the FFPE samples, among which 1119 could correspond to metabolites from the HMDB database. 905 distinct metabolites were found in both the frozen samples and the FFPE samples.

Metabolomic markers of the gliomagenesis pathway: IDH-mutant versus IDH-wild-type tumors

As a first step, metabolomic profiles were compared between *IDH*-mutant gliomas and *IDH*-wildtype gliomas regardless of histological subtype and grade, in order to identify which metabolites were preferentially produced in each molecular pathway.

Thirty-five metabolites were selected (supplementary file 1), among which, only hydroxyglutaric acid ($[M-H2O+H]^+$ adduct) was selected in more than 85% of the 100 resamples (selection probability of 100%). The mean AUC of the 100 models was 96.8%.

As shown in figure 1, the mean values of hydroxyglutaric acid (2HG) $[M-H2O+H]^+$ and $[M-H]^+$ adducts were significantly higher in the *IDH*-mutant samples than in the *IDH*-wt samples, both in the frozen sample cohort ($p < 0.001$) and in the FFPE sample cohort ($p = 0.0025$ and 0.0044 respectively).

The identification of 2HG was validated both in MS (mass difference of 1.55 ppm) and MS/MS (78.1 % match rate) (supplementary file 2).

The accuracy with which *IDH* mutation could be predicted using 2HG ($[M-H2O+H]^+$ adduct) values was measured using logistic regression and **4-fold cross validation on the frozen tissue data**. This analysis revealed a mean AUC of 96.6%, a mean accuracy of 94.3%, a mean sensitivity of 98.0% and a mean specificity of 90.4%. When using the 2HG $[M-H]^+$ adduct, the same analysis revealed an AUC of 95.5%, an accuracy of 94.3%, a sensitivity of 95.8% and a specificity of 92.2% (Figure 2).

When training a logistic regression model using the 2HG ($[M-H2O+H]^+$ adduct) values of the frozen samples and **predicting the *IDH* mutational status of the FFPE samples**, AUC was 82.6%, accuracy was 74.6%, sensitivity was 70.6% and specificity was 80.4% (Figure 3A). When using the 2HG $[M-H]^+$ adduct, AUC was 80.7%, accuracy was 74.6%, sensitivity was 74.6% and specificity was 74.6%.

Metabolomic markers of grade: grade IV versus lower grades

As a second step, grade IV tumors were compared with lower grades in order to determine which metabolites were associated with the highest grade.

A first analysis included both *IDH* mutant and *IDH* wild-type tumors:

Thirty-eight metabolites were selected (supplementary file 3A), among which 3 metabolites were selected in more than 85% of the 100 resamples (supplementary file 3A): Aminoadipic acid (AAA) (selection probability of 100%), a peak of unidentified positive ion at m/z 256.0929+ (selection probability of 100%) and guanidinoacetic acid (GAA) (selection probability of 99%). The mean AUC of the 100 models was 97%.

As shown in figure 3, among the 3 most selected metabolites, only AAA and GAA had similar differential expressions between grade IV samples and lower grade glioma samples both in frozen samples and in FFPE samples. Furthermore, AAA and GAA identification was validated in MS (mass differences of 0.89 ppm and 3.64 ppm respectively) and in MS/MS (match rates of 70.6 and 69% respectively) but the peak of unidentified ion at m/z 256.0929+ could not be identified (Supplementary file 2). For this reason, only AAA and GAA were kept for further model training and testing.

The accuracy with which grade IV gliomas could be identified using AAA and GAA values was measured using logistic regression and 4-fold cross validation on frozen tissue data. This analysis revealed a mean AUC of 94.3%, a mean accuracy of 92.0%, a mean sensitivity of 94.6% and a mean specificity of 90.8% (Figure 4A).

When training a logistic regression model on frozen samples and predicting the grade of FFPE samples, AUC was 84.7%, accuracy was 74.6%, sensitivity was 78.1% and specificity was 73.4% (Figure 3B).

In an additional step, *IDH*-mutant and *IDH*-wt astrocytomas were separately analyzed:

Concerning *IDH*-mutant astrocytomas:

The metabolite selection step led to the selection of 21 metabolites (supplementary file 3B), among which 2 metabolites were selected in more than 85% of the 100 resamples: AAA (selection probability

of 100%) and the previously described peak of unidentified positive ion at m/z value 256.0929+ (selection probability of 100%). The mean AUC of the 100 models was 93.4%.

Only AAA had similar differential expression between *IDH*-mutant grade IV astrocytomas and *IDH*-mutant low-grade astrocytomas in both the frozen samples and the FFPE samples. This difference was statistically significant in frozen samples ($p = 0.0084$) but not in FFPE samples ($p = 0.1$).

The accuracy with which *IDH*-mutant glioblastomas could be identified using the values of AAA was measured using logistic regression and 4-fold cross validation on frozen tissue data. This analysis revealed a mean AUC of 90.2%, a mean accuracy of 90.4%, a mean sensitivity of 91.7% and a mean specificity of 89.6% (Figure 4B).

When training a logistic regression model on frozen samples and predicting the grade of FFPE samples, AUC was 86.8%, accuracy was 85.0%, sensitivity was 75.0% and specificity was 88.6% (Figure 4B).

Concerning *IDH*-wt astrocytomas:

The metabolite selection step led to the selection of 24 metabolites (supplementary file 3C), among which 8 metabolites were selected in more than 85% of the 100 resamples: GAA (selection probability of 100%) and the previously described unidentified positive ion at m/z 256.0929+ (selection probability of 100%) and 6 other unidentified ions (selection probability between 86 and 97%). The mean AUC of the 100 models was 89.8%. GAA had the highest weight (supplementary file 3C), was the only metabolite with similar differential expression between glioblastomas and *IDH*-wt lower grade astrocytomas both in frozen samples and in FFPE samples (figure 3) and was the only metabolite that could be validated by MS and MS/MS analysis.

The accuracy with which glioblastomas and *IDH*-wt lower grade astrocytomas could be identified using GAA values was measured using logistic regression and 4-fold cross validation on frozen tissue data. This analysis revealed a mean AUC of 92.1%, a mean accuracy of 89.7 %, a mean sensitivity of 88.8% and a mean specificity of 90.8% (Figure 4C).

When training a logistic regression model on frozen samples and predicting the grade of FFPE samples, AUC was 92.3%, accuracy was 88.2%, sensitivity was 81.3% and specificity was 91.4% (Figure 4C).

Discussion

Metabolomic data reflect the consequences of phenotypic and genetic variations in tumors. Diffuse gliomas are a phenotypically and genetically heterogeneous group of tumors, for which WHO classification was revised in 2016. This revision imposes updates to metabolomic data that have been previously identified according to new classification criteria. In our study, we performed an untargeted metabolomic analysis on a large cohort of diffuse gliomas classified according to the WHO 2016 classification, identified a small selection of metabolites of interest in frozen samples and showed that these metabolites could be detected and used to perform classification in FFPE samples.

While writing this study, the 5th edition of the WHO Classification was published [19]. We will discuss our results in regard to the revised 2016 version, and then in light of the 2021 WHO classification.

To the best of our knowledge, this is the first study using a non-targeted metabolomic approach on FFPE samples of gliomas. Only one previous study confirmed that 2HG was overexpressed in FFPE *IDH*-mutant gliomas, as expected, using a targeted analysis [13]. As anticipated, we found metabolomic analysis of FFPE samples to be less accurate than that of frozen samples. This was most likely due, at least in part, to the fixation and dehydration steps required in FFPE sample preparation, which may alter metabolites. However, this analysis was sufficient to detect and quantify hydroxyglutaric acid, AAA and GAA in FFPE samples as well as to use these results to classify these tumor samples with specificities and sensitivities over 70%. Moreover, working with FFPE samples could offer important advantages. Indeed, FFPE samples are the most available material in routine practice. Hence, working with them can offer the advantage of larger cohorts, leading to more statistical power, but also the inclusion of rare entities for which frozen tissues are not always available, such as low-grade diffuse *IDH*-wt gliomas.

When performing metabolomic studies, statistical procedure plays a very important role in the analysis of the results. Many machine learning methods have been used in this setting [20-22]. In our study we chose to use the widely known LASSO penalized logistic regression along with bootstrapping to select a small number of metabolites of interest. LASSO penalization reduces the risk of overfitting by assuming sparse solutions leading to simpler models based on a few key features (in this case, metabolites). The use of bootstrapping enabled the estimation of selection probability for each metabolite. Focusing on a small number of key metabolites also offers the advantage of fewer false discoveries. Indeed, metabolomic data include a high number of redundant and noisy features and LASSO penalized logistic regression is robust to these unwanted features [18]. In our study, this method led to the selection of a very small number of relevant metabolites in the frozen samples, for which the predictive values could be confirmed in the FFPE samples, confirming the absence of overfitting.

Using an untargeted metabolomic analysis offers both the possibility to verify the presence of known biomarkers and the possibility to identify novel biomarkers. In this study, we found both known and novel biomarkers revealing the gliomagenesis pathway and tumor grade.

Concerning the gliomagenesis pathway (i.e. *IDH*-mutant status), hydroxyglutaric acid was, as expected, significantly overexpressed in *IDH*-mutant gliomas. As only one study has previously shown [13], we confirm that this biomarker shows a high sensitivity and a high specificity in both frozen and FFPE samples. Interestingly, we found that the $[M\text{-}H2O+H]^+$ adduct of hydroxyglutaric acid had a slightly better diagnostic power than the $[M-H]^+$ adduct in FFPE samples. It is unclear whether this adduct was generated by the LC-MS methodology or if it was previously present in the samples. Nevertheless, this adduct could be an alternative to the $[M-H]^+$ adduct for the detection of *IDH* mutation if, for instance, performing a single LC-MS analysis in positive ionization mode.

Concerning grade, we found two main metabolites of interest: aminoadipic acid (AAA) and guanidinoacetic acid (GAA). Overall, we found that a logistic regression model based on these two metabolites could predict the grade glioma samples with an accuracy of 93.1% for frozen samples and

74% for FFPE samples. Interestingly, AAA was more relevant in *IDH*-mutant astrocytomas, for which it could be used to predict the grade with an accuracy of 95.2% in frozen samples and 85% in FFPE samples, whereas GAA was more relevant in *IDH*-wt astrocytomas, for which it could be used to predict the grade with an accuracy of 97.2% in frozen samples and 88.2% in FFPE samples.

AAA is a product of the Lysine catabolism [23]. The association between AAA and glioblastomas has been poorly described as of yet, but AAA is an emergent metabolite in the medical metabolomic literature [24]. In the study of Locasale et al. [25], AAA was significantly overexpressed in the cerebrospinal fluid of patients suffering from high-grade gliomas as compared to controls without any malignancy. Using ^{1}H NMR spectroscopy, Rosi et al. [26] showed that AAA was a marker of glioblastoma-stem cell aggressiveness, such cells being correlated with a dismal prognosis. As in our study, Bjorblom *et al*. [27] found that AAA was increased in *IDH*-mutant glioblastomas as compared to lower grade *IDH*-mutant gliomas. AAA was also associated with high grade gliomas in the study of Gorynska *et al*. [28]. A study on 95 frozen samples of prostatic adenocarcinomas [29] showed that AAA was the only metabolite of which the amount was significantly associated with the TNM status and Gleason grade. These studies as well as our own support the hypothesis that AAA is correlated with tumor aggressiveness. Effects of AAA accumulation in the brain are not yet understood. AAA is known to be a gliotoxin [30] and has been shown to induce oxidative stress in the brain of rats [31].

To our knowledge, this is the first study to report an association between Guanidinoacetic acid (GAA) levels and tumor grade. GAA, also known as glycocyamine or betacyamine, is the direct precursor of creatine. Here, we found that GAA was significantly increased in grade IV gliomas as compared to lower grades. One known cause of elevated GAA is Guanidinoacetic Acid Methyl Transferase (GAMT) deficiency [32], a genetic disorder transmitted in an autosomal recessive fashion and linked to mutations on chromosome 19p13.3. In the case of GAMT deficiency, elevated GAA levels are generally associated with lower creatine levels and lower creatine/creatinine ratios. Hence, if our observation of elevated GAA in grade IV gliomas is a marker of de novo GAMT deficiency, we would

expect grade IV gliomas to have significantly lower creatine levels and creatine/creatinine ratios. However, although we found significantly lower creatine levels in frozen grade IV glioma samples (p=0.04), this was not the case in FFPE samples (grade IV gliomas had higher creatine levels with p=0.073). Furthermore, creatine/creatinine ratios were significantly increased in grade IV glioma frozen samples (p=0.0005) but non-significantly decreased in grade IV glioma FFPE samples (p=0.054). Our observation of significantly increased GAA levels in grade IV gliomas as compared to lower grades will need to be verified and explained in future studies.

The 2021 WHO classification of CNS tumors has brought some changes in the grading of both *IDH*-mutant and *IDH*-wt DG. Indeed, *IDH*-mutant GBM are now called *IDH*-mutant grade 4 astrocytomas, and an *IDH*-mutant astrocytoma harboring a homozygous deletion of *CDKN2A* is now systematically classified as grade 4, regardless of its morphology. In our study, none of the low-grade *IDH*-mutant astrocytomas had a homozygous deletion of *CDKN2A*, thus their grades remained unchanged in the 2021 WHO classification. However, all the low-grade *IDH*-wt astrocytomas in our study should now be considered grade 4, i.e. *IDH*-wt glioblastomas. Our results suggest that, although these morphologically low-grade tumors are now assimilated to glioblastomas, they are metabolically distinct, at least in the expression of GAA.

Hence, we have shown that 2HG, AAA and GAA are metabolites of interest for the classification of glial tumors. It would be of particular interest to detect these metabolites *in vivo* using proton magnetic resonance spectroscopy (1H-MRS). As 2HG is a known biomarker of *IDH* mutation, several previous studies have been performed to evaluate its detectability with 1H-MRS [33, 34]. However significant technical limitations remain and 2-hydroxyglutarate 1H-MRS is not widely used at present. To our knowledge, only very few studies have evaluated the detectability and usefulness of AAA using 1H-MRS in glial tumors. Righi et al. have recently identified AAA in a glioblastoma *in vivo* using 1H-MRS [35]. However, they report that AAA and 2HG cannot be distinguished by one dimensional 1H-MRS alone but only with two dimensional techniques such as TOCSY. We did not find any studies concerning GAA 1H-MRS in glial tumors. However, Ensenauer et al. have detected GAA using 1H-MRS in the brain

of a patient with GAMT deficiency [36]. Even though GAA levels may be higher in the brains of patients with GAMT deficiency than in glial tumors, it would be of interest to study the possible association between GAA levels as assessed by 1H-MRS and glial tumor grades.

We recognize some limitations to our study, most of which are associated with the non-targeted metabolomic approach. Indeed, although these techniques have raised interest in the last decade, the assessments using these techniques show different limitations. The initial raw data include a very high number of peaks of distinct retention times and mass/charge ratios. Working with such high dimensional dataset induces a risk of false discovery. To limit this risk as much as possible, we used different methods:

1. We filtered the initial raw data, removing peaks of lower intensities. This step aims to denoise the data but could also discard low-concentrated metabolites of interest.
2. We used a constrained statistical method, the LASSO penalized logistic regression, robust to correlated features and to overfitting,
3. We systematically verified the identification of the selected ions, by comparing the obtained mass/charge ratios with mass spectrometry databases, and by comparing the corresponding MS/MS spectrums to spectrum databases.
4. We verified the results obtained on the frozen sample cohort on the FFPE sample cohort, which we used as a validation dataset.

Technical variations represent another limitation of the untargeted metabolomic techniques because they can induce batch effects. To reduce putative batch effect in our study, we performed the LC-MS/MS analysis of the frozen samples in a single continuous batch, with a minimal number of columns. Afterwards, we performed the analysis of FFPE samples in a second batch. Here, we focus our study on metabolites found in both frozen and FFPE samples therefore the metabolites were identified in two different batches of LC-MS/MS analyses.

Finally, even though it would have been interesting to explore possible metabolomic variations associated with histological sub-types of *IDH*-mutant gliomas, we could not perform this analysis because we did not include enough oligodendrogliomas to be statistically relevant.

In conclusion, we have shown that the $[M-H]^+$ and $[M-H2O+H]^+$ adducts of 2-hydroxyglutarate could be detected and used to determine *IDH* mutational status both in frozen and FFPE samples, and that aminoadipic acid and guanidinoacetic acid could be detected and used to evaluate the grade of astrocytomas both in frozen and FFPE samples.

Furthermore, untargeted metabolomic studies can be performed on FFPE samples of adult diffuse gliomas and could represent promising tools to further understand the heterogeneity of these tumors and to develop related supervised classification methods.

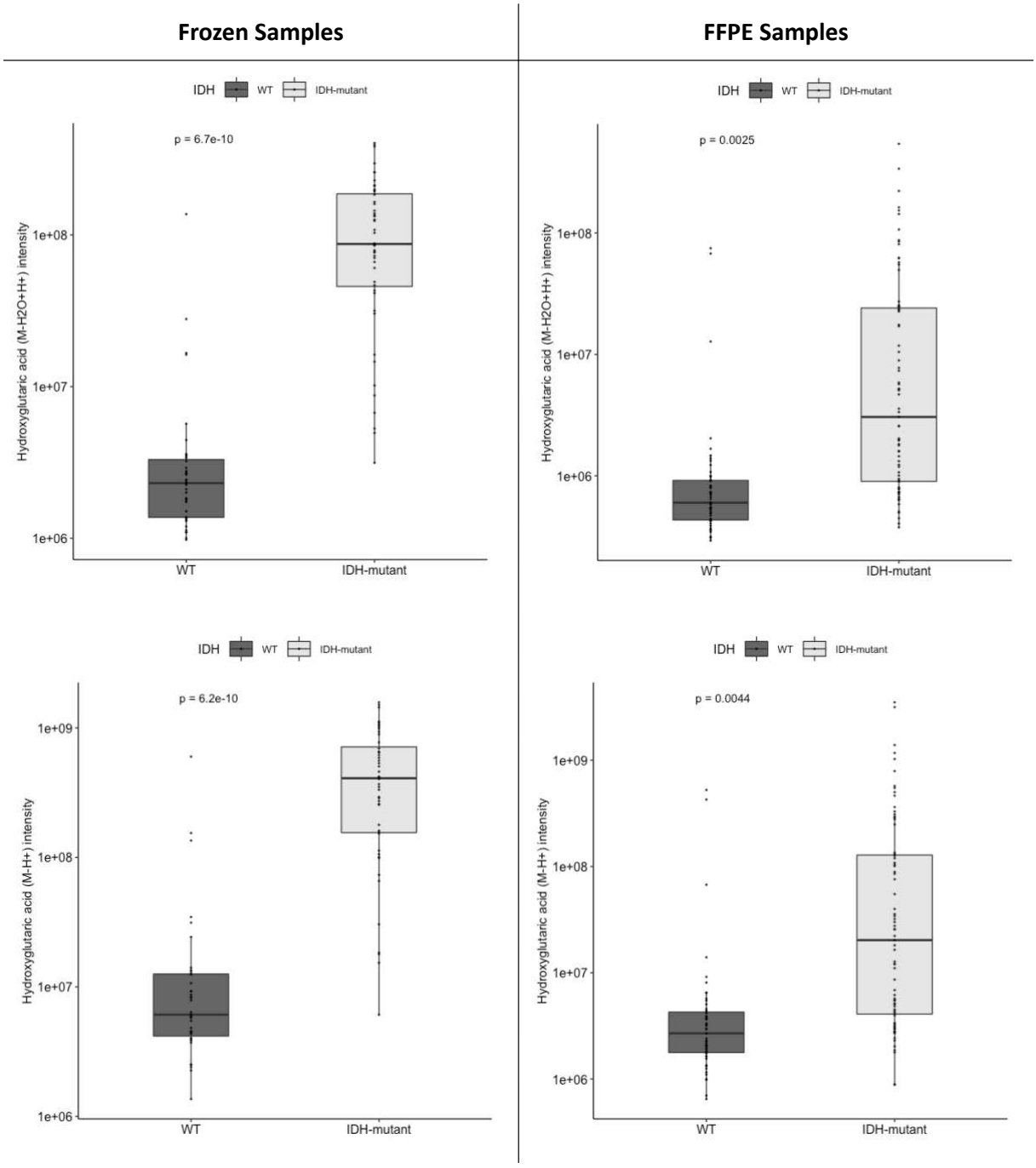

Figure 1.

Title: Boxplots of hydroxyglutaric acid levels in *IDH-wt* and *IDH-mutant* glial tumors in frozen and Formalin-Fixed and Paraffin-Embedded (FFPE) tumor samples.

Legend:

Top. Hydroxyglutaric [M-H2O+H+] adduct levels in *IDH-wt* and *IDH-mutant* glial tumors.

Bottom. Hydroxyglutaric [M-H+] adduct levels in *IDH-wt* and *IDH-mutant* glial tumors.

Scale is logarithmic. p-values are given for a student T-test.

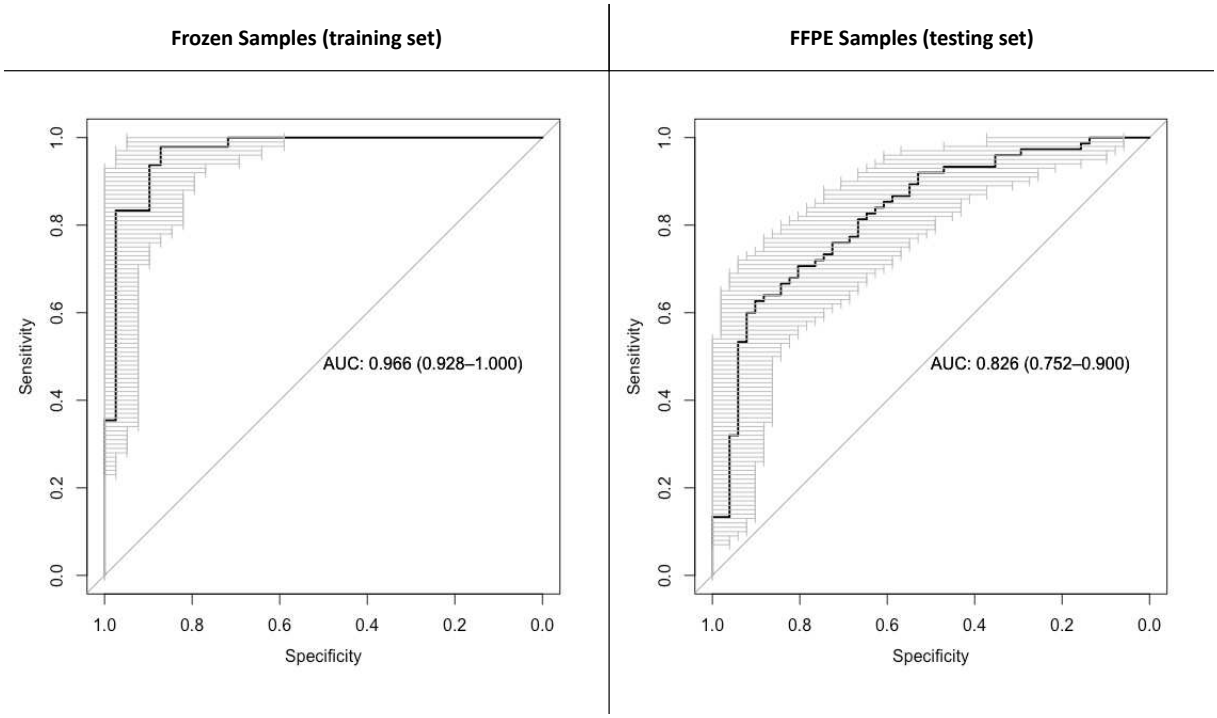

Figure 2.

Title: Receiver Operating Characteristic (ROC) curves for the logistic regression model predicting IDH mutational status based on Hydroxyglutaric [M-H2O+H+]

Legend :

ROC curves concerning the training set (frozen samples on the left) and the testing set (Formalin-Fixed and Paraffin-Embedded samples on the right).

Figure 3.

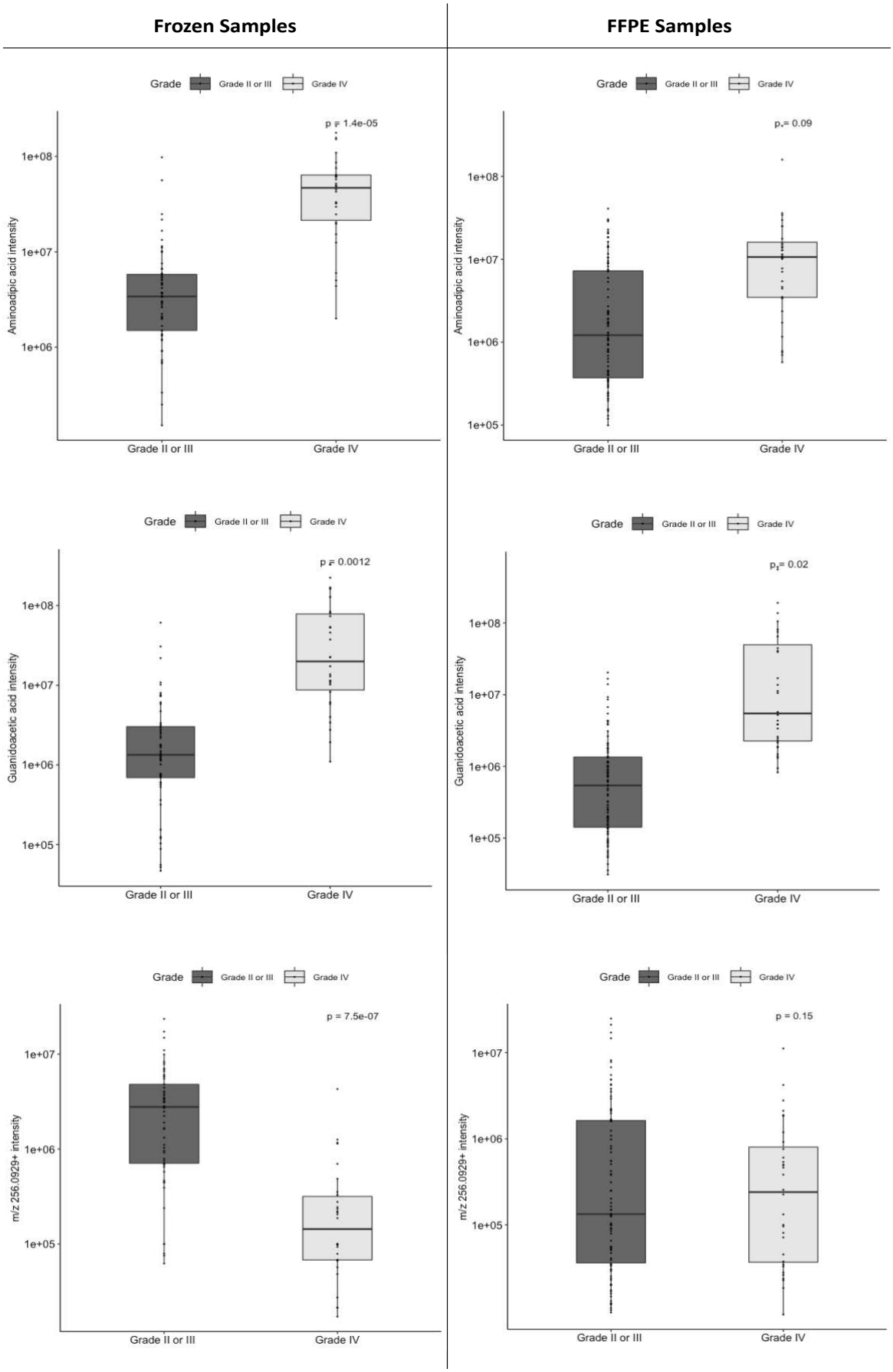
Frozen Samples
FFPE Samples
Grade
Grade II or III
Grade IV
p = 1.4e-05
p = 0.09
Aminoadipic acid intensity
1e+08
1e+07
1e+06
1e+05
Grade II or III
Grade IV
p = 0.0012
p = 0.02
Guanidoacetic acid intensity
p = 7.5e-07
p = 0.15
m/z 256.0929+ intensity

Figure 4.

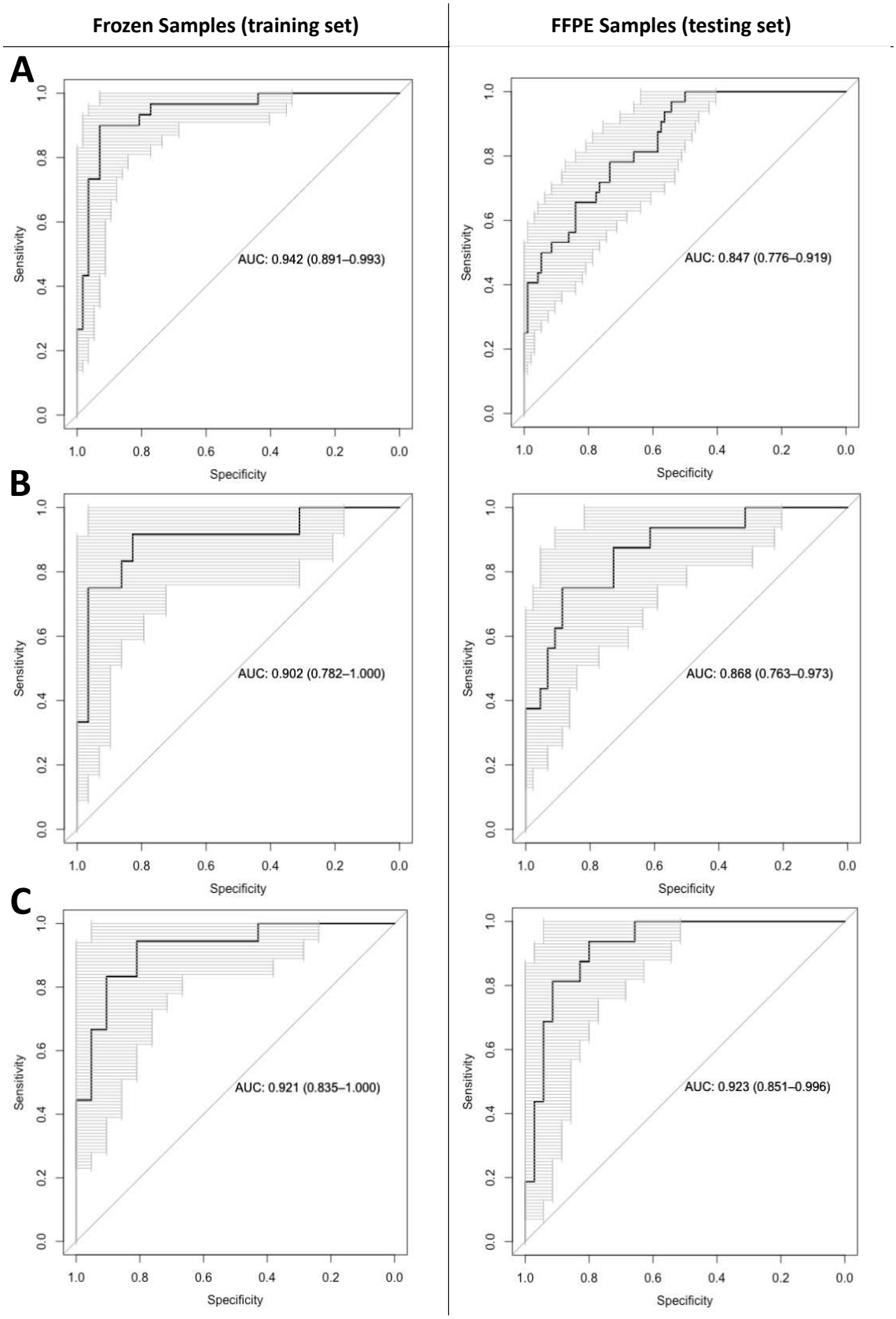
Frozen Samples (training set)
FFPE Samples (testing set)
A
B
C
Sensitivity
Specificity
AUC: 0.942 (0.891–0.993)
AUC: 0.847 (0.776–0.919)
AUC: 0.902 (0.782–1.000)
AUC: 0.868 (0.763–0.973)
AUC: 0.921 (0.835–1.000)
AUC: 0.923 (0.851–0.996)

Figure 3.

Title: Boxplots of relevant metabolites in lower grade (II or III) and high grade (IV) glial tumors in frozen and Formalin-Fixed and Paraffin-Embedded (FFPE) tumor samples.

Legend:

Top. Aminoadipic acid (AAA) levels in *IDH-wt* and *IDH-mutant* glial tumors.

Middle. Guanidinoacetic acid (GAA) levels in lower grade (II or III) and high grade (IV) glial tumors.

Bottom. Levels of an unidentified peak of m/z value 256.0929+ in lower grade (II or III) and high grade (IV) glial tumors.

Scale is logarithmic. p-values are given for a student T-test.

Figure 4.

Title: Receiver Operating Characteristic (ROC) curves concerning predictive models for the classification of glial tumors.

Legend:

All models are trained on frozen samples, ROC curves are then generated concerning the training set (frozen samples on the left) and the testing set (Formalin-Fixed and Paraffin-Embedded samples on the right).

A. ROC curves for the logistic regression model predicting histological grade for all glial tumors, regardless of IDH mutational status, based on Aminoadipic acid (AAA) and Guanidinoacetic acid (GAA) levels.

B. ROC curves for the logistic regression model predicting histological grade for IDH-mutant astrocytomas, based on AAA levels.

C. ROC curves for the logistic regression model predicting histological grade IDH-wt tumors, based on GAA levels.

References

1. Chinnaiyan, P., et al., *The metabolomic signature of malignant glioma reflects accelerated anabolic metabolism.* Cancer Res, 2012. **72**(22): p. 5878-88.
2. Reitman, Z.J., et al., *Profiling the effects of isocitrate dehydrogenase 1 and 2 mutations on the cellular metabolome.* Proc Natl Acad Sci U S A, 2011. **108**(8): p. 3270-5.
3. Zhou, L., et al., *Integrated Metabolomics and Lipidomics Analyses Reveal Metabolic Reprogramming in Human Glioma with IDH1 Mutation.* J Proteome Res, 2019. **18**(3): p. 960-969.
4. Cacciatore, S., et al., *Metabolic Profiling in Formalin-Fixed and Paraffin-Embedded Prostate Cancer Tissues.* Mol Cancer Res, 2017. **15**(4): p. 439-447.
5. Kelly, A.D., et al., *Metabolomic profiling from formalin-fixed, paraffin-embedded tumor tissue using targeted LC/MS/MS: application in sarcoma.* PLoS One, 2011. **6**(10): p. e25357.
6. Ostrom, Q.T., et al., *CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the United States in 2006-2010.* Neuro Oncol, 2013. **15 Suppl 2**: p. ii1-56.
7. Louis, D.N., Ohgaki, H., Wiestler, O.D. and Cavenee, W.K., *WHO Classification of Tumors of the Central Nervous System*. revised 4th ed. 2016, Lyon: IARC.
8. Brat, D.J., et al., *cIMPACT-NOW update 3: recommended diagnostic criteria for "Diffuse astrocytic glioma, IDH-wildtype, with molecular features of glioblastoma, WHO grade IV".* Acta Neuropathol, 2018. **136**(5): p. 805-810.
9. Sahm, F., et al., *Farewell to oligoastrocytoma: in situ molecular genetics favor classification as either oligodendroglioma or astrocytoma.* Acta Neuropathol, 2014. **128**(4): p. 551-9.
10. Dang, L., et al., *Cancer-associated IDH1 mutations produce 2-hydroxyglutarate.* Nature, 2009. **462**(7274): p. 739-44.
11. Moren, L., et al., *Metabolomic Screening of Tumor Tissue and Serum in Glioma Patients Reveals Diagnostic and Prognostic Information.* Metabolites, 2015. **5**(3): p. 502-20.

12. Lee, J.E., et al., *Metabolic profiling of human gliomas assessed with NMR.* J Clin Neurosci, 2019. **68**: p. 275-280.

13. Sahm, F., et al., *Detection of 2-hydroxyglutarate in formalin-fixed paraffin-embedded glioma specimens by gas chromatography/mass spectrometry.* Brain Pathol, 2012. **22**(1): p. 26-31.

14. Pluskal, T., et al., *MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.* BMC Bioinformatics, 2010. **11**: p. 395.

15. Myers, O.D., et al., *One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks.* Anal Chem, 2017. **89**(17): p. 8696-8703.

16. Wishart, D.S., et al., *HMDB 3.0--The Human Metabolome Database in 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p. D801-7.

17. Montenegro-Burke, J.R., C. Guijas, and G. Siuzdak, *METLIN: A Tandem Mass Spectral Library of Standards.* Methods Mol Biol, 2020. **2104**: p. 149-163.

18. Bujak, R., et al., *PLS-Based and Regularization-Based Methods for the Selection of Relevant Variables in Non-targeted Metabolomics Data.* Front Mol Biosci, 2016. **3**: p. 35.

19. WHO Classification of Tumours Editorial Board, *World Health Organization Classification of Tumours of the Central Nervous System*. 5th ed. 2021, Lyon: IARC.

20. Heinemann, J., et al., *Application of support vector machines to metabolomics experiments with limited replicates.* Metabolomics, 2014. **10**: p. 1121-8.

21. Bartel, J., J. Krumsiek, and F.J. Theis, *Statistical methods for the analysis of high-throughput metabolomics data.* Comput Struct Biotechnol J, 2013. **4**: p. e201301009.

22. Broadhurst, D.I. and A.D. Kelly, *Statistical strategies for avoiding false discoveries in metabolomics and related experiments.* Metabolomics, 2006. **2**: p. 171-96.

23. Hallen, A., J.F. Jamie, and A.J. Cooper, *Lysine metabolism in mammalian brain: an update on the importance of recent discoveries.* Amino Acids, 2013. **45**(6): p. 1249-72.

24. Bellance, N., et al., *Oncosecretomics coupled to bioenergetics identifies alpha-amino adipic acid, isoleucine and GABA as potential biomarkers of cancer: Differential expression of c-Myc, Oct1 and KLF4 coordinates metabolic changes.* Biochim Biophys Acta, 2012. **1817**(11): p. 2060-71.

25. Locasale, J.W., et al., *Metabolomics of human cerebrospinal fluid identifies signatures of malignant glioma.* Mol Cell Proteomics, 2012. **11**(6): p. M111 014688.

26. Rosi, A., et al., *(1) H NMR spectroscopy of glioblastoma stem-like cells identifies alpha-aminoadipate as a marker of tumor aggressiveness.* NMR Biomed, 2015. **28**(3): p. 317-26.

27. Bjorkblom, B., et al., *Distinct metabolic hallmarks of WHO classified adult glioma subtypes.* Neuro Oncol, 2022.

28. Gorynska, P.Z., et al., *Metabolomic Phenotyping of Gliomas: What Can We Get with Simplified Protocol for Intact Tissue Analysis?* Cancers (Basel), 2022. **14**(2).

29. Jung, K., et al., *Tissue metabolite profiling identifies differentiating and prognostic biomarkers for prostate carcinoma.* Int J Cancer, 2013. **133**(12): p. 2914-24.

30. McBean, G.J., *Inhibition of the glutamate transporter and glial enzymes in rat striatum by the gliotoxin, alpha aminoadipate.* Br J Pharmacol, 1994. **113**(2): p. 536-40.

31. da Silva, J.C., et al., *alpha-Ketoadipic Acid and alpha-Aminoadipic Acid Cause Disturbance of Glutamatergic Neurotransmission and Induction of Oxidative Stress In Vitro in Brain of Adolescent Rats.* Neurotox Res, 2017. **32**(2): p. 276-290.

32. Caldeira Araujo, H., et al., *Guanidinoacetate methyltransferase deficiency identified in adults and a child with mental retardation.* Am J Med Genet A, 2005. **133A**(2): p. 122-7.

33. Choi, C., et al., *2-hydroxyglutarate detection by magnetic resonance spectroscopy in IDH-mutated patients with gliomas.* Nat Med, 2012. **18**(4): p. 624-9.

34. Kim, H., et al., *In-Vivo Proton Magnetic Resonance Spectroscopy of 2-Hydroxyglutarate in Isocitrate Dehydrogenase-Mutated Gliomas: A Technical Review for Neuroradiologists.* Korean J Radiol, 2016. **17**(5): p. 620-32.

35. Righi, V., et al., *A metabolomic data fusion approach to support gliomas grading.* NMR Biomed, 2020. **33**(3): p. e4234.

36. Ensenauer, R., et al., *Guanidinoacetate methyltransferase deficiency: differences of creatine uptake in human brain and muscle.* Mol Genet Metab, 2004. **82**(3): p. 208-13.

VI. CONCLUSIONS ET PERSPECTIVES

Ainsi, dans le cadre de cette thèse nous avons pu évaluer de nouvelles méthodes de classification supervisée dans le cadre d'études de métabolomique. Nous avons pu tirer les conclusions suivantes :

- Il existe des alternatives intéressantes à la PLS-DA pour l'analyse statistique des données de métabolomique. Celles-ci peuvent être préférées pour leur meilleures performances et/ou pour leurs meilleures interprétabilités.

- Il est impératif d'inclure une étape de sélection des métabolites lors de l'analyse statistique pour limiter le risque d'overfitting du modèle et la surinterprétation des résultats.

- Cette étape de sélection des métabolites doit être adaptée à l'objectif de l'étude. Dans le cadre d'une étude exploratoire, visant à mieux comprendre un phénomène biologique, il parait logique de sélectionner un nombre assez large de variables. Si l'objectif est de créer un modèle prédictif, applicable en pratique courante, la sélection du nombre le plus restreint possible de métabolites semble plus pertinent, pour optimiser l'interprétabilité et la reproductibilité du modèle, pour ce faire la méthode BOLASSO nous semble pertinente.

- Il parait pertinent d'inclure une mesure de l'impact de chaque métabolite dans le modèle. Pour ce faire les poids des métabolites représente la mesure la plus directe et donc la plus intuitive.

- Il semble également pertinent d'associer un score de probabilité aux prédictions dans le cadre d'une application clinique. Celui-ci permet à l'utilisateur de mesurer la fiabilité de la prédiction et de confronter ce résultat aux autres informations disponibles.

Les méthodes pour lesquelles nous avons aidé au développement au cours de cette thèse ont été implémentées en langage python et sont librement disponibles sur internet (https://github.com/CyprienGille/Supervised-Autoencoder et https://github.com/tirolab/PD-CR). Cependant, certains utilisateurs peuvent préférer un format plus intuitif, tels que celui

proposé pour les méthodes de Metaboanalyst.ca [57]. Nous développons actuellement une interface plus intuitive pour une meilleure distribution de ces méthodes.

Ensuite, dans le cadre d'une application dédiée à une utilisation en pratique courante, nous nous sommes intéressés aux modifications métabolomiques associées à la mutation IDH et au grade histologique des tumeurs gliales de l'adulte. Nous avons travaillé sur échantillons tumoraux congelés et fixés en paraffine.

Nous avons ainsi montré qu'il était possible de réaliser des analyses de métabolomique non ciblée sur des échantillons fixés en paraffine et d'en tirer des résultats pertinents. Travailler sur échantillons fixés comporte de nombreux avantages liés à la faciliter de conservation de ces échantillons par rapport à des échantillons congelés. Cela facilite le transport d'échantillons d'un centre à l'autre et cela permet de travailler sur les très nombreux échantillons fixés conservés en tumorothèque, incluant notamment des tumeurs rares, difficilement accessibles sous formes congelées.

Nous avons pu développer deux modèles fiables, simples et reproductibles, permettant de prédire le grade histologique et le statut mutationnel IDH à partir d'une analyse métabolomique sur échantillons congelés ou fixés en paraffine. Ces modèles sont basés sur quelques métabolites pertinents (2HG, AAA et GAA). Ces métabolites ont un rationnel biologique établi pour le 2-hydroxy-glutarate et plausible pour les acides amino-adipique et guanidino-acétique. L'analyse ciblée de nouveaux échantillons permettrait de valider ces modèles et de les utiliser en pratique courante en complément des techniques déjà disponibles. De plus, l'exploration des phénomènes biologiques à l'origine de l'association entre le grade de malignité des tumeurs gliales et les acides amino-adipique et guanidino-acétique pourrait permettre de mieux comprendre leur cancérogénèse.

VII. BIBLIOGRAPHIE

1 Oliver SG, Winson MK, Kell DB, *et al.* Systematic functional analysis of the yeast genome. *Trends in Biotechnology* 1998;**16**:373–8. doi:10.1016/S0167-7799(98)01214-1

2 Fiehn O. Metabolomics – the link between genotypes and phenotypes. *Plant Mol Biol* 2002;**48**:155–71. doi:10.1023/A:1013713905833

3 Brown SA. Circadian Metabolism: From Mechanisms to Metabolomics and Medicine. *Trends in Endocrinology & Metabolism* 2016;**27**:415–26. doi:10.1016/j.tem.2016.03.015

4 Walsh MC, Nugent A, Brennan L, *et al.* Understanding the metabolome – challenges for metabolomics. *Nutrition Bulletin* 2008;**33**:316–23. doi:10.1111/j.1467-3010.2008.00732.x

5 Pluskal T, Castillo S, Villar-Briones A, *et al.* MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 2010;**11**:395. doi:10.1186/1471-2105-11-395

6 Schiffman C, Petrick L, Perttula K, *et al.* Filtering procedures for untargeted LC-MS metabolomics data. *BMC Bioinformatics* 2019;**20**:334. doi:10.1186/s12859-019-2871-9

7 Systematic Feature Filtering in Exploratory Metabolomics: Application toward Biomarker Discovery | Analytical Chemistry. https://pubs.acs.org/doi/10.1021/acs.analchem.1c00816 (accessed 12 Dec 2022).

8 Wishart DS, Tzur D, Knox C, *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res* 2007;**35**:D521–6. doi:10.1093/nar/gkl923

9 Montenegro-Burke JR, Guijas C, Siuzdak G. METLIN: A Tandem Mass Spectral Library of Standards. *Methods Mol Biol* 2020;**2104**:149–63. doi:10.1007/978-1-0716-0239-3_9

10 Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2006;**2**:171–96. doi:10.1007/s11306-006-0037-z

11 Nuzzo R. Scientific method: Statistical errors. *Nature* 2014;**506**:150–2. doi:10.1038/506150a

12 Armstrong RA. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics* 2014;**34**:502–8. doi:10.1111/opo.12131

13 Nilsson NJ 1933-2019. *The quest for artificial intelligence : a history of ideas and achievements*. Cambridge ; New York : Cambridge University Press, 2010. 2010. https://search.library.wisc.edu/catalog/9910113233802121

14 Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press 2016.

15 Perou CM, Sørlie T, Eisen MB, *et al.* Molecular portraits of human breast tumours. *Nature* 2000;**406**:747–52. doi:10.1038/35021093

16 Gal J, Bailleux C, Chardin D, *et al.* Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer. *Comput Struct Biotechnol J* 2020;**18**:1509–24. doi:10.1016/j.csbj.2020.05.021

17 Silver D, Hubert T, Schrittwieser J, *et al.* Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. 2017. doi:10.48550/arXiv.1712.01815

18 Jonsson A. Deep Reinforcement Learning in Medicine. *KDD* 2019;**5**:18–22. doi:10.1159/000492670

19 *Dynamic Programming*. 2010. https://press.princeton.edu/books/paperback/9780691146683/dynamic-programming (accessed 10 Dec 2022).

20 Belkin M, Hsu D, Ma S, *et al.* Reconciling modern machine learning practice and the bias-variance trade-off. *Proc Natl Acad Sci USA* 2019;**116**:15849–54. doi:10.1073/pnas.1903070116

21 Kantz ED, Tiwari S, Watrous JD, *et al.* Deep Neural Networks for Classification of LC-MS Spectral Peaks. *Anal Chem* 2019;**91**:12407–13. doi:10.1021/acs.analchem.9b02983

22 Zhang X, Lin T, Xu J, *et al.* DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis. *Analytica Chimica Acta* 2019;**1058**:48–57. doi:10.1016/j.aca.2019.01.002

23 Chardin D, Pourcher T, Gal J, *et al.* Métabolomique et imagerie TEP-FDG des cancers du sein. *Médecine Nucléaire* 2021;**45**:4–12. doi:10.1016/j.mednuc.2020.03.002

24 Liebal UW, Phan ANT, Sudhakar M, *et al.* Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites* 2020;**10**:243. doi:10.3390/metabo10060243

25 Pérez-Enciso M, Tenenhaus M. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Hum Genet* 2003;**112**:581–92. doi:10.1007/s00439-003-0921-9

26 Ringnér M. What is principal component analysis? *Nat Biotechnol* 2008;**26**:303–4. doi:10.1038/nbt0308-303

27 Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* 2017;**33**:2776–8. doi:10.1093/bioinformatics/btx299

28 Kim S, Kang D, Huo Z, *et al.* Meta-analytic principal component analysis in integrative omics application. *Bioinformatics* 2018;**34**:1321–8. doi:10.1093/bioinformatics/btx765

29 Rodríguez-Pérez R, Fernández L, Marco S. Overoptimism in cross-validation when using partial least squares-discriminant analysis for omics data: a systematic study. *Anal Bioanal Chem* 2018;**410**:5981–92. doi:10.1007/s00216-018-1217-1

30 Bujak R, Daghir-Wojtkowiak E, Kaliszan R, *et al.* PLS-Based and Regularization-Based Methods for the Selection of Relevant Variables in Non-targeted Metabolomics Data. *Front Mol Biosci* 2016;**3**:35. doi:10.3389/fmolb.2016.00035

31 Ruiz-Perez D, Guan H, Madhivanan P, *et al.* So you think you can PLS-DA? *BMC Bioinformatics* 2020;**21**:2. doi:10.1186/s12859-019-3310-7

32 Chong I-G, Jun C-H. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* 2005;**78**:103–12. doi:10.1016/j.chemolab.2004.12.011

33 Ding L, Yuan L, Sun Y, *et al.* Rapid Assessment of Exercise State through Athlete's Urine Using Temperature-Dependent NIRS Technology. *J Anal Methods Chem* 2020;**2020**:8828213. doi:10.1155/2020/8828213

34 Tran TN, Afanador NL, Buydens LMC, *et al.* Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). *Chemometrics and Intelligent Laboratory Systems* 2014;**Complete**:153–60. doi:10.1016/j.chemolab.2014.08.005

35 Belmonte-Sánchez JR, Romero-González R, Arrebola FJ, *et al.* An Innovative Metabolomic Approach for Golden Rum Classification Combining Ultrahigh-Performance Liquid Chromatography-Orbitrap Mass Spectrometry and Chemometric Strategies. *J Agric Food Chem* 2019;**67**:1302–11. doi:10.1021/acs.jafc.8b05622

36 Breiman L. *Classification and Regression Trees*. Routledge editions 1984. doi:10.1201/9781315139470

37 Quinlan JR. Induction of decision trees. *Mach Learn* 1986;**1**:81–106. doi:10.1007/BF00116251

38 Loh W-Y, Shih Y-S. Split Selection Methods for Classification Trees. *Statistica Sinica* 1997;**7**:815–40.

39 Friedman JH. Multivariate Adaptive Regression Splines. *The Annals of Statistics* 1991;**19**:1–67. doi:10.1214/aos/1176347963

40 Breiman L. Random Forests. *Machine Learning* 2001;**45**:5–32. doi:10.1023/A:1010933404324

41 Boser BE, Guyon IM, Vapnik VN. A Training Algorithm for Optimal Margin Classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York, NY, USA: : Association for Computing Machinery 1992. 144–52. doi:10.1145/130385.130401

42 Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97. doi:10.1007/BF00994018

43 Holland JH. *Adaptation in natural and artificial systems*. Michigan, USA: : University of Michigan Press 1975.

44 McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 1943;**5**:115–33. doi:10.1007/BF02478259

45 Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 1958;**65**:386–408. doi:10.1037/h0042519

46 Lecun Y. A Theoretical Framework for Back-Propagation. 2001.

47 Henglin M, Claggett BL, Antonelli J, *et al.* Quantitative Comparison of Statistical Methods for Analyzing Human Metabolomics Data. *Metabolites* 2022;**12**:519. doi:10.3390/metabo12060519

48 Leclercq M, Vittrant B, Martin-Magniette ML, *et al.* Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. *Front Genet* 2019;**10**. doi:10.3389/fgene.2019.00452

49 Mendez KM, Reinke SN, Broadhurst DI. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics* 2019;**15**:150. doi:10.1007/s11306-019-1612-4

50 Barlaud M, Chambolle A, Caillau J-B. Robust supervised classification and feature selection using a primal-dual method. *arXiv:190201600 [cs, stat]* Published Online First: 5 February 2019.http://arxiv.org/abs/1902.01600 (accessed 26 May 2019).

51 Chambolle A, Pock T. On the ergodic convergence rates of a first-order primal–dual algorithm. *Math Program* 2016;**159**:253–87. doi:10.1007/s10107-015-0957-3

52 Barlaud M, Chambolle A, Caillau J-B. Robust supervised classification and feature selection using a primal-dual method. *ArXiv* 2019.

53 Condat L. Fast Projection onto the Simplex and the l1 Ball. *Mathematical Programming, Series A* 2016;**158**:575–85. doi:10.1007/s10107-015-0946-6

54 Trong TN, Mehtonen J, González G, *et al.* Semisupervised Generative Autoencoder for Single-Cell Data. *J Comput Biol* 2020;**27**:1190–203. doi:10.1089/cmb.2019.0337

55 Kingma DP, Welling M. Auto-Encoding Variational Bayes. 2014. doi:10.48550/arXiv.1312.6114

56 Barlaud M, Guyard F. A Non-Parametric Supervised Autoencoder for discriminative and generative modeling. In: *ICASSP*. Toronto, Canada: 2022. https://hal.archives-ouvertes.fr/hal-02937643 (accessed 11 Dec 2022).

57 Pang Z, Zhou G, Ewald J, *et al.* Using MetaboAnalyst 5.0 for LC–HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. *Nat Protoc* 2022;**17**:1735–61. doi:10.1038/s41596-022-00710-w

58 Louis DN, Perry A, Reifenberger G, *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* 2016;**131**:803–20. doi:10.1007/s00401-016-1545-1

59 Louis DN, Perry A, Wesseling P, *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol* 2021;**23**:1231–51. doi:10.1093/neuonc/noab106

60 Bach F. Bolasso: model consistent Lasso estimation through the bootstrap. 2008.http://arxiv.org/abs/0804.1302 (accessed 16 Oct 2022).

ANNEXE 1

Computational and Structural Biotechnology Journal 18 (2020) 1509–1524

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer

Jocelyn Gal [a,1,*], Caroline Bailleux [b,1], David Chardin [c,d,1], Thierry Pourcher [d], Julia Gilhodes [e], Lun Jing [d], Jean-Marie Guigonis [d], Jean-Marc Ferrero [b], Gerard Milano [f], Baharia Mograbi [g], Patrick Brest [g], Yann Chateau [a], Olivier Humbert [c,d], Emmanuel Chamorey [a]

[a] University Côte d'Azur, Epidemiology and Biostatistics Department, Centre Antoine Lacassagne, Nice F-06189, France
[b] University Côte d'Azur, Medical Oncology Department Centre Antoine Lacassagne, Nice F-06189, France
[c] University Côte d'Azur, Nuclear Medicine Department, Centre Antoine Lacassagne, Nice F-06189, France
[d] University Côte d'Azur, Commissariat à l'Energie Atomique, Institut de Biosciences et Biotechnologies d'Aix-Marseille, Laboratory Transporters in Imaging and Radiotherapy in Oncology, Faculty of Medicine, Nice F-06100, France
[e] Department of Biostatistics, Institut Claudius Regaud, IUCT-O Toulouse, France
[f] University Côte d'Azur, Centre Antoine Lacassagne, Oncopharmacology Unit, Nice F-06189, France
[g] University Côte d'Azur, CNRS UMR7284, INSERM U1081, IRCAN TEAM4 Centre Antoine Lacassagne FHU-Oncoage, Nice F-06189, France

ARTICLE INFO

Article history:
Received 11 February 2020
Received in revised form 15 May 2020
Accepted 16 May 2020
Available online 3 June 2020

Keywords:
Unsupervised machine learning
Metabolomics
Breast neoplasms
Computer simulation

ABSTRACT

Genomics and transcriptomics have led to the widely-used molecular classification of breast cancer (BC). However, heterogeneous biological behaviors persist within breast cancer subtypes. Metabolomics is a rapidly-expanding field of study dedicated to cellular metabolisms affected by the environment. The aim of this study was to compare metabolomic signatures of BC obtained by 5 different unsupervised machine learning (ML) methods. Fifty-two consecutive patients with BC with an indication for adjuvant chemotherapy between 2013 and 2016 were retrospectively included. We performed metabolomic profiling of tumor resection samples using liquid chromatography-mass spectrometry. Here, four hundred and forty-nine identified metabolites were selected for further analysis. Clusters obtained using 5 unsupervised ML methods (PCA k-means, sparse k-means, spectral clustering, SIMLR and k-sparse) were compared in terms of clinical and biological characteristics. With an optimal partitioning parameter k = 3, the five methods identified three prognosis groups of patients (favorable, intermediate, unfavorable) with different clinical and biological profiles. SIMLR and K-sparse methods were the most effective techniques in terms of clustering. *In-silico* survival analysis revealed a significant difference for 5-year predicted OS between the 3 clusters. Further pathway analysis using the 449 selected metabolites showed significant differences in amino acid and glucose metabolism between BC histologic subtypes. Our results provide proof-of-concept for the use of unsupervised ML metabolomics enabling stratification and personalized management of BC patients. The design of novel computational methods incorporating ML and bioinformatics techniques should make available tools particularly suited to improving the outcome of cancer treatment and reducing cancer-related mortalities.

© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Breast cancer (BC) is the most common type of cancer in women worldwide and the second leading cause of cancer-associated deaths [1]. The treatment strategy may be guided by two classifications indicating the aggressiveness of the tumor. The anatomy-clinical classification is based on age, TNM, histological factors (histological grade, Ki-67) as well as on hormonal-receptor status and Her-2 expression. The molecular classification resulting from genomic [2], transcriptomic [3] and proteomic [4] analyses introduced the concept of luminal A, luminal B, Her-2 and basal-like BC [5–7]. This latter classification from Perou and Sorlie was assessed using unsupervised analyses [6,8]. Efforts have been made to develop multivariate prognostic models such as, AdjuvantOnline®,

* Corresponding author at: Department of Epidemiology and Biostatistics, Centre Antoine Lacassagne, University Côte d', Azur 33 avenue de Valombrose, 06189 Nice, France.
E-mail address: jocelyn.gal@nice.unicancer.fr (J. Gal).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2020.05.021
2001-0370/© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

PREDICT Tool [9,10] and multigene predictors [11,12]. The use of biomarker-based tests, including omics-based tests, has steadily increased over the last decade as a result of the need for personalized treatment strategies designed to optimize outcomes [13–18]. Several genomic prognostic markers have been described for BC such as OncotypeDX®, Prosigna®, MammaPrint®, Endopredict® Genomic grade index® and BC Index® [19]. Two markers are commercially available and are increasingly used in clinical practice (21-gene recurrence score OncotypeDX® and 70-gene prognostic signature MammaPrint®). However, heterogeneity persists in biological features within BC subtypes, thus highlighting the need to improve the taxonomy [20]. This heterogeneity may be related to specific combinations of genetic, pathological and environmental factors leading to specific metabolic alterations and interactions [21,22].

Metabolomics is a new and growing field dedicated to the study of metabolism at overall level that promises to provide new insights into disease mechanisms and drug effects. Indeed, metabolomics may offer a complementary approach to genomics and could be used to better understand the influence of the environment on tumor phenotype [23]. Two distinct approaches characterize metabolomics: a targeted approach aimed at quantifying as accurately as possible a limited number of predefined metabolites of interest [24] and an untargeted approach aimed at measuring, without any a priori, as many metabolites as possible in a sample [25,26]. As with other omics approaches, metabolomics generates high-dimensional data. The processing of these data can be done by applying supervised or unsupervised machine learning (ML) algorithms that are increasingly used for medical diagnosis and therapeutic strategy guidance [27–29]. Unsupervised ML, in which no a priori class label information is given to guide the algorithm [30], seems a suitable alternative to analyze these data and address the problem of BC heterogeneity [6]. The aim of this study was to compare metabolomic signatures of BC obtained using five different unsupervised ML methods. To evaluate the consistency of our results, the clusters obtained by unsupervised ML methods were compared with patients' clinical characteristics and identified metabolic pathways.

2. Material and methods

2.1. Patients

This is a retrospective cohort study based on data and samples from 52 patients already available in the Centre Antoine Lacassagne tumor bank and collected during routine practice between 2013 and 2016. Patient tumor characteristics were: clinical stages I to III_B biopsy-proven BC, with an indication for post-surgery adjuvant therapy. Tumor phenotypes were classified into three subtypes: triple-negative (estrogen receptor, progesterone receptor and Her-2 non-over-expressed); luminal (estrogen receptor and/or progesterone receptor positive and Her-2 non-over-expressed); Her-2 over-expressed (Her-2 over-expressed, estrogen receptor and progesterone receptor either positive or negative) [31]. After surgery, all patients were treated according to current guidelines, with sequential chemotherapy including anthracyclines (epirubicin and cyclophosphamide) and taxanes followed by radiotherapy. Patients with Her-2 over-expressed tumors were treated with trastuzumab concurrently with taxanes and continued for one year. Patients with luminal BC were then treated by endocrine therapy with tamoxifen or an aromatase inhibitor, based on menopausal status. Clinical, histological, radiological and therapeutic data were retrospectively extracted from our facility's digital records or collected by a clinical data monitor. Follow-up data were either extracted from our facility's digital records or retrieved by telephone if patients had changed facilities during surveillance. Written informed consent was obtained from all study participants. All procedures performed in this study involving tissue collection and analyses were following the ethical standards of the institutional and/or national research committee (French National Commission for Informatics and Liberties N°17003 and National Institute Health data N°1515251018).

2.2. Data-preprocessing, metabolite identification, statistical and pathway analysis

Sample collection, preparation and data-processing using MZmine [32,33] are shown in Supplementary Material S1 and Supplementary Fig. 1 Metabolites obtained from positive and negative ionization modes were combined. Only metabolites with no null values after pre-processing were selected for analysis. When a metabolite was detected in both positive and negative modes, only the mode offering the highest average intensity was considered. After these steps, 1271 metabolites were identified. To eliminate noisy data, a filtering function was applied before statistical analysis. Finally, statistical analysis was performed on 449 metabolites. The identification of metabolic pathways was performed using MetaboAnalyst database sources [34]. The impact score was determined by the relative pathway topological effect of the metabolites, and -log*(p)* was used as the enrichment score, reflecting the probability of the pathway being identified at random; the number of "hits" was the actual number of matched metabolites in the pathway. For the selection of the most relevant pathways, we applied the following criteria: Impact >0, FDR < 0.25 and $p < 0.05$ [35].

A Venn diagram (http://bioinformatics.psb.ugent.be/webtools/Venn/) was used to display all possible logical relations between the metabolites or pathways identified by the clustering methods. Differences between clusters regarding the most active metabolites were plotted using boxplots.

2.3. Clustering algorithms

Five unsupervised clustering methods were selected and compared: Principal Component Analysis (PCA) k-means, Sparse k-means, Single-cell Interpretation via Multi-kernel LeaRning (SIMLR), k-sparse and Spectral clustering. Many clustering approaches exist, among which two of the most popular are K-means and spectral clustering [36]. PCA k-means and Sparse k-means are two well established, K-means based methods frequently used in computational. SIMLR and K-sparse are two recently developed k-means based methods of particular interest for omics data. These methods use different dimension reduction steps with k-means. In order to apply these five unsupervised clustering methods, the optimal number of clusters was determined in advance using five criteria: gap [37], silhouette [38,39], Davies-Bouldin [40], Calinski-Harabasz [41] and SIMLR method [42]. PCA k-means clustering, combines PCA to reduce the number of dimensions of a dataset and the k-means method to minimize the intra-cluster variance for a chosen number of k clusters [43–45]. Spectral clustering [46,47] is based on graph theory. It consists of identifying dense regions in a multidimensional dataset, i.e. observations that can form a non-convex set but are close to each other. Sparse k-means clustering was developed in 2010 by Witten and Tibshirani [8]. This method is based on a Least Absolute Shrinkage and Selection Operator (LASSO) approach [48] and combines the LASSO approach and the k-means method which simultaneously find the clusters and select features. SIMLR clustering [42] was developed to analyze scRNA-seq data. This method searches for appropriate cell-to-cell similarity metrics to perform dimension reduction and clustering. In multiple-kernel learning frameworks, this

method may be especially beneficial for data containing no identifiable clusters. K-sparse clustering [49] is an algorithm combining dimension reduction and relevant feature selection using a constraint in L1-norm rather than a lasso-type penalty to select the features. The performance of an unsupervised clustering method is measured by its ability to partition data. Partitioning is considered optimal when it minimizes the average distance between patients within a cluster (homogeneity) and maximizes cluster distances 2 by 2 (separability). The performances of the five methods were compared using the silhouettes index (SI) [39]. The SI ranges between −1 and 1 and assesses whether a patient belongs to the "right" cluster. The closer the index is to 1, the more satisfactory the assignment of a patient to a cluster. The t-SNE method was used for data visualization [50]. Processing times were obtained on a computer using an i5 processor (3.1 GHz).

2.4. Clinical evaluation

The relevance of the discovered clusters was assessed by comparing the clinical and survival characteristics between clusters using χ^2 or Fisher's exact tests for categorical data, analysis of variance or Mann-Whitney's test for continuous variables and log-rank test for censored data. Overall survival (OS) was defined as the time between diagnosis and death due to any cause. Specific survival (SS) was determined by the time between diagnosis and death due to BC. Recurrence-Free Survival (RFS) was defined as the time between diagnosis and the first recurrence (local, regional and metastasis). Patients showing no event (death or recurrence) or lost to follow-up were censored at the date of their last contact. OS, SS, and RFS were estimated using the Kaplan-Meier method. Median follow-up with a 95% confidence interval was calculated by reverse Kaplan–Meier method. All analyses were performed with Matlab® R2018b for PCA k-means, Spectral clustering, SIMLR (https://github.com/BatzoglouLabSU/SIMLR/tree/SIMLR/MATLAB) and k-sparse clustering and R [51] using package Sparcl [52] for sparse k-means clustering. The difference between clusters regarding the most biologically significant metabolites was plotted using boxplots. For clinical and biological analyses, all p-values <0.05 (two-sided) were considered statistically significant.

2.5. Prediction for 5- and 10-year overall and specific survival

Web-based prognostication PREDICT tool (https://breast.predict.nhs.uk/tool) [9,10,53] was used to estimate predicted OS (pOS) and predicted SS (pSS) at 5 and 10 years, based on several patient and tumor characteristics. For each patient, ten characteristics were entered manually: age at diagnosis, menopausal status, estrogen receptor status, Her-2 status, Ki-67 status, tumor stage, histological grade, mode of detection, number of positive nodes and presence of micrometastases. PREDICT tool can be used to estimate expected overall survival at 5 years and 10 years in the absence of available survival data due to short follow-up. If information was missing for detection, bisphosphonate therapy or menopausal status, patients were not excluded but the "unknown" category was used. Only one patient was excluded because of missing tumor grade data. A 1000 resamples bootstrap was used to estimate the 95% confidence interval.

3. Results

3.1. Patient characteristics

Tumor and treatment features of the 52 patients were described in Table 1. Median age was 63 years (range: 37–88). The main histological type was invasive ductal carcinoma (92%), and the main tumor stages were T1 (40.5%) and T2 (46%). Twenty-four patients (46%) presented axillary lymph node invasion. Two patients (4%) were oligometastatic at diagnosis. Forty-three percent of patients had histological grade II tumors and 47% had grade III tumors. Half of the patients had negative hormone receptor status (48%) and 24% of patients had Her-2 over-expression. Median follow–up was 48.5 months (95%CI [43–54.5]). Twenty-one patients presented a recurrence: 4 local recurrences (7.5%), 6 regional recurrences (11.5%) and 11 metastatic recurrences (21%). Three-year OS was 90% [82–99], 3-year SS was 92% [85–100] and 3-year RFS was 82% [72–93] (Supplementary Fig. 2). Median OS, SS, and RFS were not reached.

Table 1
Patients' demographics and treatment characteristics.

Clinical characteristic	No. of patients	%
Age (median min – max)	63.2 (37–88)	
Histology type		
Invasive ductal carcinoma	48	92
Invasive lobular carcinoma	3	6
Microinvasive carcinoma	1	2
Tumor stage		
T1	21	40.5
T2	24	46
T3	7	13.5
Axillary lymph node status		
N0	28	54
N+	24	46
Metastasis		
M0	50	96
M1	2	4
Histological grade		
I	5	10
II	22	43
III	24	47
Hormonal receptors status*		
Negative	25	48
Positive	27	52
Her-2 status		
Non-over-expressed	40	74
Over-expressed	12	24
Triple-negative status		
No	37	71
Yes	15	29
Tumor phenotype		
Her2	12	23
Luminal	25	48
Triple-Negative	15	29
Adjuvant Chemotherapy		
No	13	25
Yes	39	75
Adjuvant Radiotherapy		
No	9	17
Yes	43	83
Adjuvant Hormonotherapy		
No	24	46
Yes	28	54

* Oestrogen and/or progesterone.

3.2. Clustering results

3.2.1. Estimated number of clusters

Using four methods (Gap statistic, Calinski-Harabasz, Silhouette and SIMLR criterion), the optimal number of clusters was equal to three (k = 3) (Supplementary Fig. 3). Only for Davies-Bouldin criterion, the optimal number of clusters was equal to four (k = 4). It

seems reasonable, therefore, to conclude that the optimal number of clusters is equal to 3.

3.2.2. Patient distribution

Three clusters were identified with each of the five clustering methods, (Fig. 1). In terms of processing times, PCA k-means was the fastest and K-sparse was the longest (Supplementary Table 1). SIMLR and k-sparse methods were the most discriminants with an average silhouette value of 0.85 and 0.91, respectively (Fig. 2). Seventy-three percent of patients (38/52) were ranked in the same clusters by the five methods, 17.5% of patients (9/52) were classified in the same clusters by 4 methods and 9.5% of patients (5/52) were classified in the same clusters by 3 methods.

3.2.3. Comparison of clinical characteristics between clusters

As shown in Table 2, the 5 methods revealed significant inter-cluster differences. Patients in cluster 3 had mainly unfavorable prognostic factors: tumor stage T2/T3, histological grade III, high mitotic score and triple-negative phenotype. In contrast, patients in cluster 1 had mainly favorable prognosis factors: tumor stage T1, histological grade I/II, lower mitotic score and luminal phenotype, whereas patients in cluster 2 constitute an intermediate group presenting both good and poor prognostic factors. Clusters defined by PCA k-means were significantly different for 5 characteristics: tumor stage, mitosis, tumor phenotype, Her-2 status and luminal. Clusters defined by Spectral Clustering were significantly different for 6 characteristics: tumor stage, histological grade, mitosis, Ki67, tumor phenotype and luminal. Clusters defined by Sparse k-means were significantly different for 4 characteristics: histological grade, tumor phenotype, Her-2 status and luminal. Clusters defined by SIMLR were significantly different for 6 characteristics: tumor stage, histological grade, mitosis, Ki67, tumor phenotype and luminal. Clusters defined by K-Sparse were significantly different for 6 characteristics: tumor stage, histological grade, mitosis, Ki67, tumor phenotype and luminal. From a strictly clinical point of view, Spectral clustering, SIMLR and K-sparse are the 3 most discriminating methods. Indeed, for these 3 methods, six prognostic factors (tumor stage, histological grade, mitosis score, Ki-67, tumor phenotype and luminal) were distributed significantly different between the 3 clusters.

3.2.4. Comparison of survival and predicted survival between clusters

None of the methods created clusters showing significant differences for OS, SS or RFS. Analysis of patients' simulated survival data

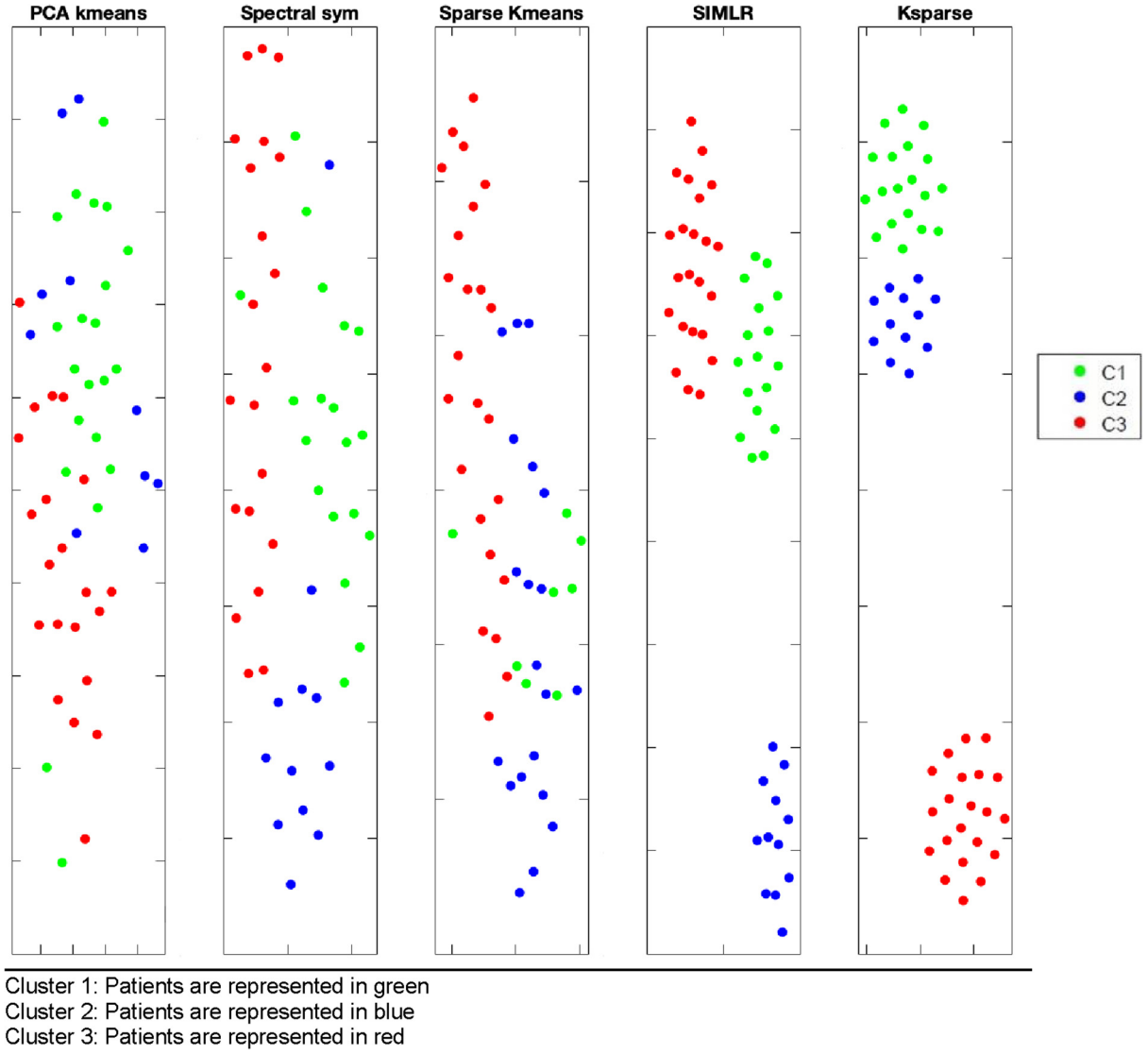

Fig. 1. Visualization of each cluster by clustering method using T-sne.

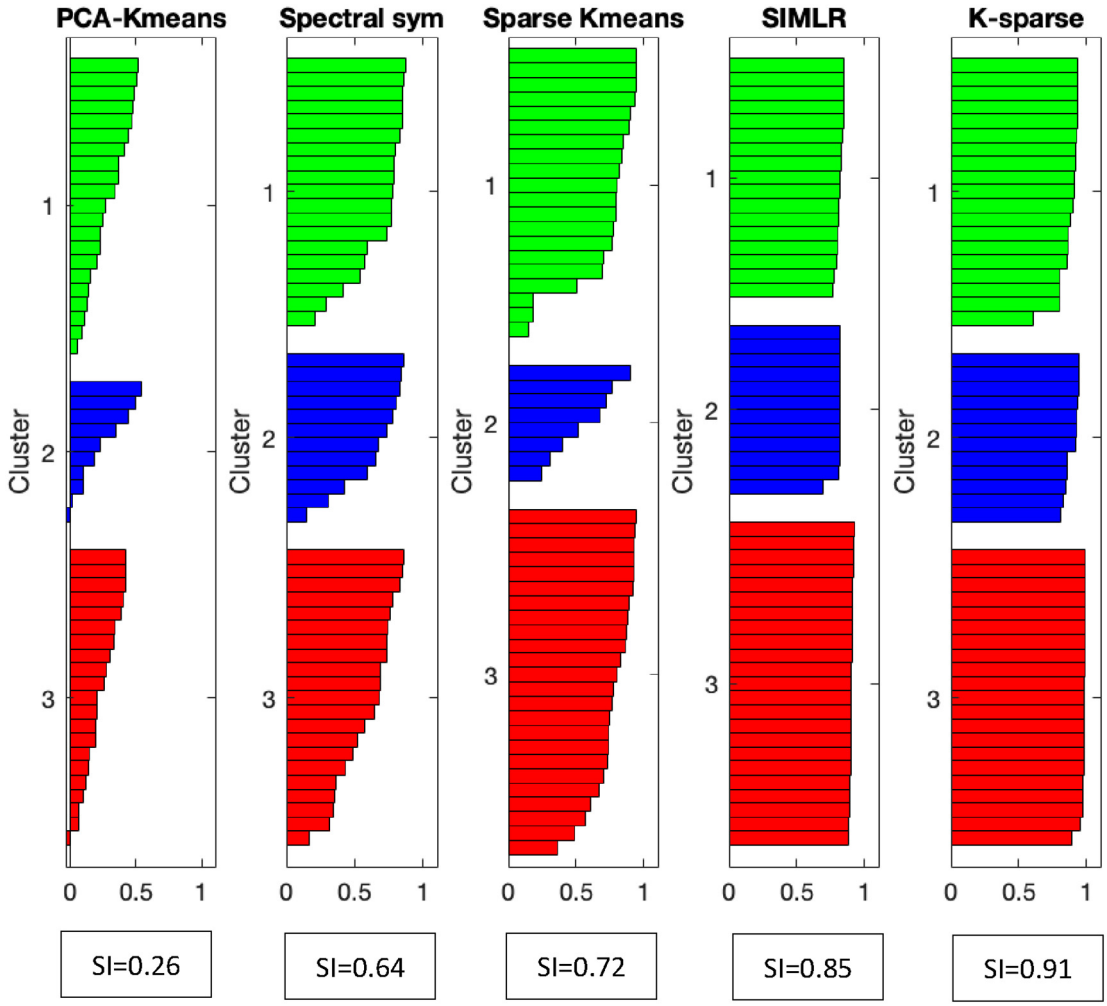

Fig. 2. Silhouette value (SI) representation for each patient by clustering method.

using PREDICT tool are presented in Table 3 and show a predicted survival gradient for clusters obtained with the 5 methods for OS and SS. There were significant differences for 5-year pOS between clusters obtained with K-sparse (p = 0.021), Sparse K-means (p = 0.049), Spectral and clustering (p = 0.021). The five methods showed a significant difference for 5-year pSS between clusters. In terms of 10-year pOS, there were no significant differences between clusters obtained by any of the 5 methods. In contrast, for 10-year pSS, the 5 methods showed significant differences between clusters. Patients in cluster 3 clearly showed the poorest predicted survival.

3.2.5. Comparison of the most impactful metabolites according to the five methods

To relate the impact of 449 metabolites to cluster construction, we ranked these metabolites extracted from each of the five methods based on their functional contributions to outputs. With this approach, we classified the relative impact of metabolites on cluster construction and on the identification of metabolic signatures. The highest-ranked metabolites were those that provided relevant information to the signature versus those that provided redundant information or no information. Among a total of 449 metabolites, 116 (26%) were selected by K-sparse clustering and 69 (15%) by Sparse K-means clustering. As for the three other methods, which don't select sparse features, the number of metabolites remained equal to 449. The 50 most effective metabolites identified by the five methods are presented in Supplementary Table 2. Furthermore, a comparison of the top 50 metabolites in each of the 5 methods is presented using a Venn diagram (Fig. 3). Two metabolites were shared by the 5 methods (Creatine, L-Proline), 9 were shared by 4 methods (Betaine, Glutathione, Humulinic Acid A, Isoleucyl-Methionine, L-Carnitine, L-Methionine, L-Phenylalanine Triethanolamine, Alnustone), 28 were shared by 3 methods and 38 were shared by 2 methods (Table 4).

3.2.6. Comparison between 5 methods of identified metabolic pathways

For a better understanding of metabolic dysregulation among BC subtypes, pathway analysis was performed. Identification of all the metabolic pathways highlighted by each of the 5 methods as shown in Supplementary Table 3. The most relevant pathways for each of the 5 methods are shown in Table 5. Sparse K-means identified only one statistically significant pathways, "cysteine and methionine metabolism", involved in amino acid metabolism. K-Sparse identified 3 different pathways: "glycerolipid metabolism", "Starch and sucrose metabolism" involved in carbohydrates metabolic pathway and "Aminoacyl-tRNA biosynthesis" involved in translation pathway. Spectral clustering identified 17 pathways, the 3 most important being "Glycine, serine and threonine metabolism", "Alanine, aspartate and glutamate metabolism" and "Histidine metabolism and glutathione metabolism" involved in amino acid metabolic pathway. PCA K-

Table 2
Clinical comparison of 52 patients between clusters.

Clinical characteristic	PCA-K-means				Spectral Clustering				Sparse K-means				SIMLR				K-Sparse			
	C1 (N = 21)	C2 (N = 10)	C3 (N = 21)	P-value	C2 (N = 19)	C1 (N = 12)	C3 (N = 21)	P-value	C1 (N = 24)	C2 (N = 8)	C3 (N = 20)	P-value	C1 (N = 17)	C2 (N = 12)	C3 (N = 23)	P-value	C1 (N = 19)	C2 (N = 12)	C3 (N = 21)	P-value
Age [a]	62.7 (15.2)	64.8(16)	62.9(15)	0.93	64.8 (14.3)	62.5 (16.5)	62 (15.3)	0.8	64.1(15)	60.5 (17.2)	63 (14.9)	0.85	64.3 (14.1)	64.9 (16.1)	61.4 (15.6)	0.755	64.8 (14.3)	62.5 (16.5)	62(15.3)	0.827
Histology type				1				0.392				0.106				0.752				0.392
Ductal carcinoma	19(90.5)	10(1 0 0)	19(90.5)		17(89.5)	11(91.7)	20(95.2)		21(87.5)	7(87.5)	20(1 0 0)		15(88.2)	12(1 0 0)	21(91.3)		17(89.5)	11(91.7)	20(95.2)	
Lobular carcinoma	2(9.5)	0(0)	1(4.8)		2(10.5)	1(8.3)	0(0)		3(12.5)	0(0)	0(0)		2(11.8)	0(0)	1(4.3)		2(10.5)	1(8.3)	0(0)	
Microinvasive carcinoma	0(0)	0(0)	1(4.8)		0(0)	0(0)	1(4.8)		0(0)	1(12.5)	0(0)		0(0)	0(0)	1(4.3)		0(0)	0(0)	1(4.8)	
Tumor stage				0.005				**0.018**				0.063				**0.045**				**0.018**
T1	14(66.7)	3(30)	4(19)		12(63.2)	5(41.7)	4(19)		14(58.3)	2(25)	5(25)		10(58.8)	6(50)	5(21.7)		12(63.2)	5(41.7)	4(19)	
T2/T3	7(33.3)	7(70)	17(81)		7(36.8)	7(58.3)	17(81)		10(41.7)	6(75)	15(75)		7(41.2)	6(50)	18(78.3)		7(36.8)	7(58.3)	17(81)	
Axillary lymph node				0.162				0.075				0.526				0.387				0.075
N0	14(66.7)	6(60)	8(38.1)		14(73.7)	6(50)	8(38.1)		15(62.5)	4(50)	9(45)		11(64.7)	7(58.3)	10(43.5)		14(73.7)	6(50)	8(38.1)	
N+	7(33.3)	4(40)	13(61.9)		5(26.3)	6(50)	13(61.9)		9(37.5)	4(50)	11(55)		6(35.3)	5(41.7)	13(56.5)		5(26.3)	6(50)	13(61.9)	
Metastasis				0.667				1				1				0.497				1
M0	21(1 0 0)	10(1 0 0)	19(90.5)		18(94.7)	12(1 0 0)	20(95.2)		23(96)	8(1 0 0)	19(95)		17(1 0 0)	12(1 0 0)	21(86.9)		18(94.7)	12(1 0 0)	20(95.2)	
M1	0(40)	0(0)	2(9.5)		1(5.3)	0(0%)	1(4.8)		1(4)	0(0%)	1(5)		0(0%)	0(0%)	2(13.1)		1(5.3)	0(0)	1(50)	
Histological grade				0.109				**0.025**				**0.008**				**0.007**				**0.025**
I/II	13(61.9)	7(70)	7(35)		12(63.2)	9(75)	6(30)		15(62.5)	5(71.4)	7(35)		11(64.7)	9(75)	7(31.8)		12(63.2)	9(75)	6(30)	
III	8(38.1)	3(30)	13(75)		7(36.8)	3(25)	14(70)		9(37.5)	2(28.6)	13(65)		6(35.3)	3(25)	15(68.2)		7(36.8)	3(25)	14(70)	
Mitosis				**0.024**				**0.016**				0.133				**0.005**				**0.016**
1	11(52.4)	4(40)	2(10)		10 (52.6)	5 (41.7)	2 (10)		11 (45.8)	2 (28.6)	4 (20)		10 (58.8)	5 (41.7)	2 (9.1)		10 (52.6)	5 (41.7)	2 (10)	
2	3(14.3)	4(40)	7(35)		3 (15.8)	5 (41.7)	6 (30)		4 (16.7)	4 (57.1)	6 (30)		2 (11.8)	5 (41.7)	7 (31.8)		3 (15.8)	5 (41.7)	6 (30)	
3	7(33.3)	2(20)	11(55)		6 (31.6)	2 (16.7)	10 (60)		9 (37.5)	1 (14.3)	10 (50)		5 (29.4)	2 (16.7)	13 (59.1)		6 (31.6)	2 (16.7)	12 (60)	
Ki67 [a]	25 (5,100)	27.5 (10,90)	60 (10,90)	0.066	41.1 (30.6)	33(22.6)	58.8 (27.2)	**0.027**	30 (19.2, 80)	35 (23.8, 45)	60 (28.8, 90)	0.196	38 (31)	32.8 (22.7)	59.7 (25.9)	**0.009**	41.1 (30.6)	33 (22.6)	58.8 (27.2)	**0.027**
Tumour phenotype				**0.024**				**0.012**				**0.006**				**0.018**				**0.012**
Her-2 over-expressed	1(4.8)	4(40)	7(33.3)		1(5.3)	4(33.3)	7(33.3)		2(8.3)	4(50)	6(30)		1(5.9)	4(33.3)	7(30.4)		1(5.3)	4(33.3)	7(33.3)	
Luminal	14(66.7)	5(50)	6(28.6)		13(68.4)	7(58.3)	5(23.8)		16(66.7)	4(50)	5(25)		12(70.6)	7(58.3)	6(26.1)		13(68.4)	7(58.3)	5(23.8)	
Triple-Negative	6(28.6)	1(10)	8(38.1)		5(26.3)	1(8.3)	9(42.9)		6(25)	0(0)	9(45)		4(23.5)	1(8.3)	10(43.5)		5(26.3)	1(8.3)	9(42.9)	
Hormonal receptors status				0.178				0.075				0.112				0.071				0.075
Negative	7(33.3)	5(50)	13(61.9)		6(31.6)	5(41.7)	14(66.7)		8(33.3)	4(50)	13(65)		5(29.4)	5(41.7)	15(65.2)		6(31.6)	5(41.7)	14(66.7)	
Positive	14(66.7)	5(50)	7(38.1)		13(68.4)	7(58.3)	7(33.3)		16(66.7)	4(50)	7(35)		12(70.6)	7(58.3)	8(34.8)		13(68.4)	7(58.3)	7(33.3)	
Her-2 status				**0.028**				0.061				**0.031**				0.115				0.061
Non-over-expressed	20(95.2)	6(60)	13(66.7)		18(94.7)	8(66.7)	14(66.7)		22(91.7)	4(50)	14(70)		16(94.1)	8(66.7)	16(69.6)		18(94.7)	6(66.7)	14(66.7)	
Over-expressed	1(4.8)	5(40)	6(33.3)		1(5.3)	4(33.3)	7(33.3)		2(8.3)	4(50)	6(30)		1(5.9)	4(33.3)	7(30.4)		1(5.3)	4(33.3)	7(33.3)	
Triple-Negative status				0.272				0.104				0.051				0.087				0.104
No	15(71.4)	9(90)	13(61.9)		14(73.7)	11(91.7)	12(57.1)		18(75)	8(1 0 0)	11(55)		13(76.5)	11(91.7)	13(56.5)		14(73.7)	11(91.7)	12(57.1)	
Yes	6(28.6)	1(10)	8(38.1)		5(26.3)	1(8.3)	9(42.9)		6(25)	0(0)	9(45)		4(23.5)	1(8.3)	10(43.5)		5(26.3)	1(8.3)	9(42.9)	
Luminal				**0.047**				**0.014**				**0.018**				**0.015**				**0.014**
No	7(33.3)	5(50)	15(71.4)		6(31.6)	5(41.7)	16(76.2)		8(33.3)	4(50)	15(75)		5(29.4)	5(41.7)	17(73.9)		6(31.6)	5(41.7)	16(76.2)	
Yes	14(66.7)	5(50)	6(28.6)		13(68.4)	7(58.3)	5(23.8)		16(66.7)	4(50)	5(25)		12(70.6)	7(58.3)	6(26.1)		13(68.4)	7(58.3)	5(28.8)	
Adjuvant Chemotherapy				0.52				0.423				0.459				0.459				0.423
No	7(33.3)	3(30)	4(19)		7(36.8)	2(16.7)	4(19)		6(25)	2(25)	5(25)		6(35.3)	3(25)	4(17.4)		7(36.8)	2(16.7)	4(19)	
Yes	14(85.7)	7(70)	17(81)		12(63.2)	10(83.3)	17(81)		18(75)	6(75)	1575)		11(64.7)	9(75)	19(82.6)		12(63.2)	10(83.3)	17(81)	
Adjuvant Radiotherapy				0.561				0.803				0.69				1				0.803
No	3(14.3)	3(30)	3(14.3)		3(15.8)	3(25)	3(14.3)		3(12.5)	2(25)	4(20)		3(17.6)	2(16.7)	4(17.4)		3(15.8)	3(25)	3(14.3)	
Yes	18(85.7)	7(70)	18(85.7)		16(84.2)	9(75)	18(85.7)		21(87.5)	6(75)	16(80)		14(82.4)	10(83.3)	19(82.6)		16(84.2)	9(75)	18(85.7)	

C1: cluster 1; C2: cluster 2; C3: cluster 3; [a]: mean (sd) or median (min, max).

Table 3
Comparison of prediction for overall and specific survival between clusters at 5 and 10-year.

		Predict 5-year				Predict 10-year			
		Overall Survival		Specific Survival		Overall Survival		Specific Survival	
Methods	No. of patients	% [95% CI]	P-value	% [95% CI]	P-value	% [95% CI]	P-value	% [95% CI]	P-value
K-sparse			**0.021**		**0.002**		**0.077**		**0.004**
	Cluster 1 (n = 19)	77% [67–82]		87% [80–91]		58% [48–65]		80% [73–86]	
	Cluster 2 (n = 12)	71% [57–82]		81% [69–90]		53% [38–66]		75% [60–85]	
	Cluster 3 (n = 20)	59% [47–69]		68% [60–74]		41% [29–52]		62% [53–69]	
SIMLR			0.1		**0.011**		0.241		**0.009**
	Cluster 1 (n = 17)	75% [64–82]		85% [77–91]		55% [45–64]		77% [65–84]	
	Cluster 2 (n = 12)	72% [56–82]		83% [69–91]		55% [40–67]		79% [65–87]	
	Cluster 3 (n = 22)	61% [50–70]		71% [63–77]		43% [32–53]		64% [55–70]	
Sparse K-means			**0.049**		**0.027**		0.203		**0.024**
	Cluster 1 (n = 24)	74% [64–80]		84% [76–89]		54% [43–63]		80% [73–86]	
	Cluster 2 (n = 7)	72% [58–87]		83% [70–94]		56% [37–72]		75% [60–85]	
	Cluster 3 (n = 20)	61% [49–69]		70% [61–78]		42% [32–52]		62% [53–69]	
Spectral clustering			**0.021**		**0.002**		0.077		**0.004**
	Cluster 1 (n = 19)	77% [68–83]		77% [80–91]		58% [48–65]		82% [73–86]	
	Cluster 2 (n = 12)	71% [57–81]		71% [69–90]		52% [32–64]		75% [60–85]	
	Cluster 3 (n = 20)	59% [47–68]		69% [60–76]		41% [29–52]		62% [53–69]	
PCA K-means			0.055		**0.009**		0.085		**0.008**
	Cluster 1 (n = 21)	77% [67–81]		86% [79–91]		58% [48–65]		79% [71–85]	
	Cluster 2 (n = 10)	69% [53–81]		80% [66–90]		52% [32–64]		77% [63–86]	
	Cluster 3 (n = 20)	60% [47–69]		69% [61–78]		41% [29–52]		63% [54–70]	

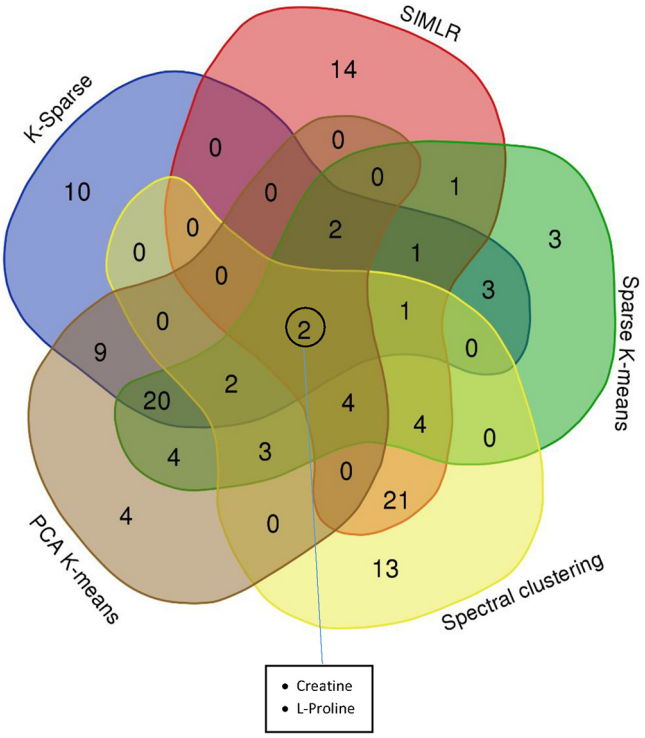

Fig. 3. Venn diagram of metabolic that were in common or unique to the five clustering methods.

means identified 10 pathways the 3 most important of which are "Alanine, aspartate and glutamate metabolism" involved in amino acid metabolic pathway, "Pyruvate metabolism" involved in carbohydrates metabolic/glucose oxidation pathway and "Citrate cycle (TCA cycle)" involved in energy metabolic pathway.

Finally, with 30 identified pathways, SIMLR is the method that identified the most metabolic pathways. Of these, the 3 most important highlighted metabolic pathways are "arginine and proline metabolism", "glycine, serine and threonine metabolism" and "alanine, aspartate and glutamate metabolism", involved in amino acid metabolic pathways. The Venn diagram (Fig. 4) shows the overlap of pathways detected by the five methods. Amino acid metabolism appeared to be the most frequently modified pathway. Enrichment and pathway analyses also showed modifications in glucose metabolism. From the biological point of view, SIMLR and spectral clustering are the two methods that identified the most relevant metabolic pathways.

3.2.7. Comparison of intensity of metabolites between the 5 methods

Among amino acid and glucose metabolisms, fourteen related metabolites were selected as potential biomarkers in BC [54–57]. As shown in Supplementary Fig. 4, the intensities of these 14 metabolites were compared between the 3 clusters for each of the 5 methods. The intensity of Uridine diphosphate (UDP) glucose, Guanine, L-Glutamine, L-Glutamic acid, L-Isoleucine, L-Proline, L-Methionine, L-Phenylalanine, Pyruvic acid, Spermine, Glutathione, Creatine, L-Carnitine and L-Acetylcarnitine were statistically significant between at least one of the clusters. The five methods agree that cluster 3 patients have low levels of Creatine, L-acetylcarnitine, L-Glutamic acid and high levels of Guanine, L-Isoleucine, L-Phenylalanine, Pyruvic acid and Spermine (Fig. 5). These metabolite levels seem to be predictive of poor prognosis [57–59].

4. Discussion

4.1. From a machine learning perspective

To the best of our knowledge, this proof-of-concept study is the first to compare different unsupervised ML methods to identify metabolomics-based prognostic signatures in BC. Analyses were performed intentionally without any prior clinical or biological assumptions. Clinical and biological interpretations were performed only after cluster identification. The objective of our study was to compare different unsupervised ML algorithms for feature selection from untargeted metabolomic data and to evaluate the capacity of these methods to select relevant features for further use in prediction models. This study did not seek to highlight significant differences but rather to assess how unsupervised methods might behave with high-dimension metabolic data and to open up new perspectives in the particularly active domain of BC

Table 4
Table indicating which metabolites are in each intersection or are unique to a certain list.

	Clustering Methods	Nbr	Metabolites
5	K-Sparse PCA K-means SIMLR Sparse K-means Spectral clustering	2	Creatine; L-Proline;
4	K-Sparse SIMLR Sparse K-means Spectral clustering	1	Triethanolamine;
	K-Sparse PCA K-means SIMLR Sparse K-means	2	L-Methionine; L-Phenylalanine
	K-Sparse PCA K-means Sparse K-means Spectral clustering	2	L-Carnitine; Betaine;
	PCA K-means SIMLR Sparse K-means Spectral clustering	4	Glutathione; Isoleucyl-Methionine; Humulinic acid A; Alnustone;
3	K-Sparse SIMLR Sparse K-means	1	Hydroxyprolyl-Valine;
	K-Sparse PCA K-means Sparse K-means	20	Aminoadipic acid; Methylmalonic acid; 1b-Furanoeudesm-4(15)-en-1-ol acetate; Glycerophosphocholine; Lidocaine; Adenosine monophosphate; 2-Methyl-3-ketovaleric acid; Liqcoumarin; p-Cresol sulfate; 2-Methylbutyroylcarnitine; Methoxsalen; Citramalic acid; Hypoxanthine; L-Acetylcarnitine; Ethyl aconitate; Guanine; L-Glutamic acid; Uridine 5′-monophosphate; N1,N12-Diacetylspermine; 5-Aminoimidazole ribonucleotide
	SIMLR Sparse K-means Spectral clustering	4	2,5-Dichloro-4-oxohex-2-enedioate; Histidinyl-Isoleucine; 3-(4-Methyl-3-pentenyl)thiophene; (−)-Epigallocatechin
	PCA K-means Sparse K-means Spectral clustering	3	L-Isoleucine; Ascorbic acid; Neurine;
2	K-Sparse Sparse K-means	3	5-Hydroxyisourate; Hexanoylcarnitine; L-Glutamine;
	K-Sparse PCA K-means	9	Creatinine; Proline; betaine; Erythronic acid; Garcinia acid; Thiolutin; 4-Chloro-1H-indole-3-acetic acid; Niacinamide 3-Dehydroxycarnitine; Dihydrothymine;
	SIMLR Spectral clustering	21	5b-Cyprinol sulfate; 2′,4-Dihydroxy-4′,6′-dimethoxychalcone; Propenoylcarnitine; 5-Hydroxyindoleacetic acid; Phaseolic acid Lisuride; 2-Bromophenol; (alpha-D-mannosyl)7-beta-D-mannosyl-diacetylchitobiosyl-L-asparagine isoform B (protein); Plastoquinone 3; 2,2,4,4,-Tetramethyl-6-(1-oxopropyl)-1,3,5-cyclohexanetrione; 1-Pyrroline; Gingerol; Prehumulinic acid; 1-Methylpyrrolo[1,2-a]pyrazine; 5-(methylthio)-2,3-Dioxopentyl phosphate; Propionic acid; Isosakuranin; Phenmetrazine; Methionine sulfoxide; Glycerol; Carboxyphosphamide
	SIMLR Sparse K-means	1	Phosphoric acid;
	PCA K-means Sparse K-means	4	l(−); L-Tyrosine; Gravelliferone; Valganciclovir;
1	K-Sparse	10	Prolylhydroxyproline; Guanidoacetic acid; Histamine; PC-M6; L-Histidine; N-Acetyl-L-aspartic acid; 3-Mercaptohexyl hexanoate; Trimethylamine N-oxide; Pantothenic acid; Flunitrazepam
	SIMLR	14	3-Hydroxy-6,8-dimethoxy-7(11)-eremophilen-12,8-olide; Glycerol tripropanoate; Alanyl-Isoleucine; 1-(2,4,6-Trimethoxyphenyl)-1,3-butanedione; 1-Oxo-1H-2-benzopyran-3-carboxaldehyde; 1,3,11-Tridecatriene-5,7,9-triyne; N-Acetyl-L-methionine; 3-Methyl sulfolene; 5-(4-Acetoxy-3-oxo-1-butynyl)-2,2′-bithiophene; Ac-Ser-Asp-Lys-Pro-OH; Cyclic AMP; Benzothiazole; (±)-2-Methylthiazolidine; 2-Methylcitric acid
	Spectral clustering	13	2,3-diketogulonate; 2,5-Furandicarboxylic acid; Pyrrolidine; Piperidine; Beta-Alanine; Aspartyl-L-proline; Erythro-5-hydroxy-L-lysinium(1 +); Acrylamide; 5-Hydroxylysine; S-Nitrosoglutathione; 2,2-dichloro-1,1-ethanediol; Valerenic acid; Dichloromethane
	Sparse K-means	3	Erinapyrone C; Ergothioneine; N-Methylethanolaminium phosphate
	PCA K-means	4	Dimethylglycine; Pipecolic acid; Methyl (9Z)-10′-oxo-6,10′-diapo-6-carotenoate; N-Desmethylvenlafaxine

phenotype predictors. We demonstrated that the K-sparse and SIMLR methods have a higher clustering performance compared with the three other popular unsupervised ML methods in detecting groups of patients with BC using metabolomic data. Interestingly, even though the spectral method is a little less clinically efficient than the k-sparse and SIMLR methods, it identified relevant metabolic pathways.

Our study suffers from various limitations, namely the relatively small number of patients and the monocentric and retrospective nature of the study. Besides, our results could not be validated on an external cohort. The clustering performances were assessed only by internal validation based on silhouette value. Indeed, we could not compare the labels obtained from our classification with the true labels to calculate the accuracy of the classification since the true labels were unknown.

Other unsupervised ML methods such as model-based clustering, bi-clustering and deep learning may be of value in this analysis and should be further explored. Yet it is worth noting that, even

Table 5
List of significant relevant pathways identified by 5 methods.

K-Sparse method							
Clusters Comparaison	Interaction metabolite	Pathway Name	Total Cmpd[a]	Match Status[b]	Raw P^c	-log(p)	Impact[d]
C1 vs C3	UDP – glucose	Starch and sucrose metabolism	50	1	0,0107	4,5388	0,1390
	UDP – glucose	Amino sugar and nucleotide sugar metabolism	88	1	0,0107	4,5388	0,0928
	UDP - glucose; Glyceric acid	Glycerolipid metabolism	32	2	0,0153	4,1831	0,0206
SIMLR method							
Clusters Comparaison	Interaction metabolite	Pathway Name	Total Cmpd	Match Status	P Value	-log(p)	Impact
C1 VS C2	Glutathione; Oxidized glutathione; Glycine; L-Glutamic acid; Pyroglutamic acid; Spermidine; Ornithine; Putrescine; Spermine; Cadaverine; Aminopropylcadaverine; Ascorbic acid	Glutathione metabolism	38	12	0	12,826	0,3628
	Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid;	Ascorbate and aldarate metabolism	45	5	0	12,469	0,1383
	L-Tryptophan; N-Acetylserotonin; 5-Hydroxyindoleacetic acid; 2-Aminomuconic acid semialdehyde; 3-Hydroxyanthranilic acid; L-Kynurenine; Acetyl-N-formyl-5-methoxykynurenamine; Isophenoxazine;	Tryptophan metabolism	79	8	0,0001	9,1233	0,2741
	5′-Methylthioadenosine; N-Formyl-L-methionine; L-Homocysteine; L-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; L-Aspartyl-4-phosphate; Pyruvic acid;	Cysteine and methionine metabolism	56	9	0,0008	7,1674	0,2509
	L-Glutamine; Phosphoribosylformylglycineamidine; Cyclic AMP; Adenosine monophosphate; Adenosine; Inosine; Adenine; Hypoxanthine; Guanine; Uric acid; 5-Hydroxyisourate; Guanosine; Adenosine diphosphate ribose; 5-Aminoimidazole ribonucleotide; Glyoxylic acid; Glycine; Adenosine 3′,5′-diphosphate;	Purine metabolism	92	17	0,0011	6,8091	0,2048
	Glyoxylic acid; Oxoglutaric acid; N-Formyl-L-methionine; Glycolic acid; Glyceric acid; Pyruvic acid;	Glyoxylate and dicarboxylate metabolism	50	6	0,0027	5,9281	0,268
	L-Glutamine; Ornithine; Citrulline; L-Arginine; L-Glutamic acid; N-Acetylornithine; L-Proline; Hydroxyproline; Guanidoacetic acid; Creatine; 4-Guanidinobutanoic acid; N2-Succinyl-L-ornithine; Putrescine; Spermidine; N-Acetylputrescine; Pyruvic acid; Glyoxylic acid; Spermine;	Arginine and proline metabolism	77	19	0,0053	5,238	0,6514
	Oxoglutaric acid; Oxalosuccinic acid; Pyruvic acid;	Citrate cycle (TCA cycle)	20	3	0,0075	4,8991	0,176
	D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,0076	4,8821	0,0394
	2-Hydroxyethanesulfonate; Pyruvic acid; 3-Sulfinoalanine;	Taurine and hypotaurine metabolism	20	3	0,0154	4,1754	0,0324
	Glyceric acid; Betaine; Guanidoacetic acid; Dimethylglycine; Glycine; Phosphoserine; L-Threonine; O-Phosphohomoserine; L-Aspartyl-4-phosphate; Creatine; Glyoxylic acid; Pyruvic acid; L-Tryptophan	Glycine, serine and threonine metabolism	48	13	0,018	4,0154	0,46986
	Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; N-Acetyl-D-Glucosamine 6-Phosphate; Uridine diphosphate-N-acetylglucosamine; Cytidine monophosphate N-acetylneuraminic acid; D-Glucose; D-Xylose	Amino sugar and nucleotide sugar metabolism	88	7	0,0187	3,9783	0,1417
	Formiminoglutamic acid; L-Glutamic acid; Urocanic acid; L-Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetol-phosphate; Oxoglutaric acid;	Histidine metabolism	44	10	0,0412	3,1903	0,3705
	Pyridoxamine; Oxoglutaric acid; 3-Hydroxy-2-methylpyridine-4,5-dicarboxylate; Pyruvic acid;	Vitamin B6 metabolism	32	4	0,0412	3,1898	0,0773
C1 VS C3	Formiminoglutamic acid; L-Glutamic acid; Urocanic acid; L-Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetol-phosphate; Oxoglutaric acid;	Histidine metabolism	44	10	0,0139	4,2752	0,3705
	Phenylpyruvic acid; L-Phenylalanine; L-Tyrosine; 3-Dehydroquinate; L-Tryptophan;	Phenylalanine, tyrosine and tryptophan biosynthesis	27	5	0,0189	3,9687	0,099
	L-Tryptophan; N-Acetylserotonin; 5-Hydroxyindoleacetic acid; 2-Aminomuconic acid semialdehyde; 3-Hydroxyanthranilic acid; L-Kynurenine; Acetyl-N-formyl-5-methoxykynurenamine; Isophenoxazine;	Tryptophan metabolism	79	8	0	16,409	0,2741
C2 VS C3	Glutathione; Oxidized glutathione; Glycine; L-Glutamic acid; Pyroglutamic acid; Spermidine; Ornithine; Putrescine; Spermine; Cadaverine; Aminopropylcadaverine; Ascorbic acid;	Glutathione metabolism	38	12	0	16,133	0,3628
	Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid	Ascorbate and aldarate metabolism	45	5	0	13,096	0,1383
	5′-Methylthioadenosine; N-Formyl-L-methionine; L-Homocysteine; L-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; L-Aspartyl-4-	Cysteine and methionine	56	9	0,0001	9,8548	0,2509

(*continued on next page*)

Table 5 (*continued*)

SIMLR method							
Clusters Comparaison	Interaction metabolite	Pathway Name	Total Cmpd	Match Status	P Value	-log(p)	Impact
	phosphate; Pyruvic acid;	metabolism					
	Phenylpyruvic acid; L-Phenylalanine; L-Tyrosine; 3-Dehydroquinate; L-Tryptophan;	Phenylalanine, tyrosine and tryptophan biosynthesis	27	5	0,0001	8,9814	0,099
	L-Histidine; L-Phenylalanine; L-Arginine; L-Glutamine; Glycine; L-Methionine; L-Lysine; L-Isoleucine; L-Threonine; L-Tryptophan; L-Tyrosine; L-Proline; L-Glutamic acid; Phosphoserine;	Aminoacyl-tRNA biosynthesis	75	14	0,0002	8,758	0,1127
	Glyoxylic acid; Oxoglutaric acid; N-Formyl-L-methionine; Glycolic acid; Glyceric acid; Pyruvic acid;	Glyoxylate and dicarboxylate metabolism	50	6	0,0004	7,7271	0,268
	L-Glutamine; Phosphoribosylformylglycineamidine; Cyclic AMP; Adenosine monophosphate; Adenosine; Inosine; Adenine; Hypoxanthine; Guanine; Uric acid; 5-Hydroxyisourate; Guanosine; Adenosine diphosphate ribose; 5-Aminoimidazole ribonucleotide; Glyoxylic acid; Glycine; Adenosine 3′,5′-diphosphate;	Purine metabolism	92	17	0,0007	7,306	0,2048
	Malonic acid; Beta-Alanine; Spermine; Spermidine; Dihydrouracil; Pantothenic acid; Uracil; L-Histidine	beta-Alanine metabolism	28	8	0,0012	6,7568	0,3577
	Uridine 5′-monophosphate; L-Glutamine; Dihydrouracil; Cytidine monophosphate; Cytidine; Cytosine; Uracil; Dihydrothymine; Uridine diphosphate glucose; Malonic acid; Ureidosuccinic acid; Beta-Alanine; Methylmalonic acid;	Pyrimidine metabolism	60	13	0,0014	6,5817	0,2756
	Pantothenic acid; Dihydrouracil; Beta-Alanine; Pyruvic acid; Adenosine 3′,5′-diphosphate; Uracil;	Pantothenate and CoA biosynthesis	27	6	0,0023	6,0879	0,2736
	L-Phenylalanine; Phenylpyruvic acid; Benzoic acid; Hippuric acid; Pyruvic acid; L-Tyrosine;	Phenylalanine metabolism	45	6	0,0072	4,9364	0,2468
	L-Glutamic acid; L-Glutamine; Oxoglutaric acid	D-Glutamine and D-glutamate metabolism	11	3	0,0124	4,39	0,139
	L-Glutamine; Ornithine; Citrulline; L-Arginine; L-Glutamic acid; N-Acetylornithine; L-Proline; Hydroxyproline; Guanidoacetic acid; Creatine; Creatinine; 4-Guanidinobutanoic acid; N2-Succinyl-L-ornithine; Putrescine; Spermidine; N-Acetylputrescine; Pyruvic acid; Glyoxylic acid; Spermine;	Arginine and proline metabolism	77	19	0,0169	4,082	0,6514
	2-Hydroxyethanesulfonate; Pyruvic acid; 3-Sulfinoalanine;	Taurine and hypotaurine metabolism	20	3	0,0215	3,8411	0,0324
	N-Acetyl-L-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; L-Glutamine; L-Glutamic acid; 2-Keto-glutaramic acid;	Alanine, aspartate and glutamate metabolism	24	7	0,0221	3,8108	0,4122
	Pyridoxamine; Oxoglutaric acid; 3-Hydroxy-2-methylpyridine-4,5-dicarboxylate; Pyruvic acid;	Vitamin B6 metabolism	32	4	0,0267	3,6235	0,0773
	Oxoglutaric acid; Oxalosuccinic acid; Pyruvic acid	Citrate cycle (TCA cycle)	20	3	0,0302	3,5015	0,176
	Glyceric acid; Betaine; Guanidoacetic acid; Dimethylglycine; Glycine; Phosphoserine; L-Threonine; O-Phosphohomoserine; L-Aspartyl-4-phosphate; Creatine; Glyoxylic acid; L-Tryptophan	Glycine, serine and threonine metabolism	48	13	0,0372	3,2914	0,4699
	Uridine diphosphate glucose; Glycerol 3-phosphate; Glycerol; Glyceric acid; Galactosylglycerol;	Glycerolipid metabolism	32	5	0,0427	3,1546	0,2162
	D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,0427	3,1536	0,0394
Sparse K-means method							
Clusters Comparaison	Interaction metabolite		Total Cmpd	Match Status	Raw p	-log(p)	Impact
C1 VS C2	L-Methionine; Glutathione	Cysteine and methionine metabolism	56	2	0.007	4.9	0.0454
C1 VS C3	L-Methionine; Glutathione;	Cysteine and methionine metabolism	56	2	0.0020	6.2	0.00454
Spectral clustering method							
Clusters Comparaison	Interaction metabolite	Pathway Name	Total Cmpd	Match Status	Raw p	-log(p)	Impact
C1 VS C3	Iminoaspartic acid; Quinolinic acid; Niacinamide; Pyruvic acid; Propionic acid;	Nicotinate and nicotinamide metabolism	44	5	0,0024	6,0206	0,0712
	Glyceric acid; Betaine; Guanidoacetic acid; Dimethylglycine; Glycine; Phosphoserine; L-Threonine; O-Phosphohomoserine; L-Aspartyl-4-phosphate; Creatine; Glyoxylic acid; L-Tryptophan	Glycine, serine and threonine metabolism	48	13	0,0040	5,5100	0,4699

Table 5 (*continued*)

Spectral clustering method							
Clusters Comparaison	Interaction metabolite	Pathway Name	Total Cmpd	Match Status	Raw p	-log(p)	Impact
	5′-Methylthioadenosine; N-Formyl-L-methionine; L-Homocysteine; L-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; L-Aspartyl-4-phosphate; Pyruvic acid;	Cysteine and methionine metabolism	56	9	0,0098	4,6232	0,2509
	Formiminoglutamic acid; L-Glutamic acid; Urocanic acid; L-Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetol-phosphate; Oxoglutaric acid;	Histidine metabolism	44	10	0,0101	4,5961	0,3705
	xoglutaric acid; Oxalosuccinic acid; Pyruvic acid;	Citrate cycle (TCA cycle)	20	3	0,0171	4,0710	0,1760
	Pyruvic acid; L-Threonine; L-Isoleucine;	Valine, leucine and isoleucine biosynthesis	27	3	0,0178	4,0277	0,0350
	D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,0210	3,8609	0,0394
	D-Glucose; Glyceric acid; Pyruvic acid;	Pentose phosphate pathway	32	3	0,0232	3,7622	0,0218
	Pyruvic acid; L-Lactic acid; D-Glucose;	Glycolysis or Gluconeogenesis	31	3	0,0249	3,6928	0,0953
	Pyruvic acid; L-Lactic acid;	Pyruvate metabolism	32	2	0,0274	3,5955	0,3201
	L-Glutamic acid; Pyruvic acid; Butyric acid; Oxoglutaric acid;	Butanoate metabolism	40	4	0,0283	3,5644	0,0852
	2-Hydroxyethanesulfonate; Pyruvic acid; 3-Sulfinoalanine;	Taurine and hypotaurine metabolism	20	3	0,0287	3,5525	0,0324
	Glyoxylic acid; Oxoglutaric acid; N-Formyl-L-methionine; Glycolic acid; Glyceric acid; Pyruvic acid;	Glyoxylate and dicarboxylate metabolism	50	6	0,0303	3,4966	0,2680
	Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid;	Ascorbate and aldarate metabolism	45	5	0,0330	3,4104	0,1383
	Epinephrine; Dopamine; L-Tyrosine; Homovanillic acid; Pyruvic acid;	Tyrosine metabolism	76	5	0,0385	3,2580	0,1750
	N-Acetyl-L-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; L-Glutamine; L-Glutamic acid; 2-Keto-glutaramic acid;	Alanine, aspartate and glutamate metabolism	24	7	0,0390	3,2431	0,4122
	Pyridoxamine; Oxoglutaric acid; 3-Hydroxy-2-methylpyridine-4,5-dicarboxylate; Pyruvic acid;	Vitamin B6 metabolism	32	4	0,0447	3,1074	0,0773
PCA K-means method							
Clusters Comparaison	Interaction metabolite	Pathway Name	Total Cmpd	Match Status	Raw p	-log(p)	Impact
C1 vs C3	Iminoaspartic acid; Quinolinic acid; Niacinamide; Pyruvic acid; Propionic acid;	Nicotinate and nicotinamide metabolism	44	5	0,003	5,9412	0,0712
	Oxoglutaric acid; Oxalosuccinic acid; Pyruvic acid;	Citrate cycle (TCA cycle)	20	3	0,011	4,4865	0,1760
	Epinephrine; Dopamine; L-Tyrosine; Homovanillic acid; Pyruvic acid;	Tyrosine metabolism	76	5	0,024	3,7311	0,1750
	Pyruvic acid; L-Lactic acid;	Pyruvate metabolism	32	2	0,043	3,1507	0,3201
	D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; Pyruvic acid;	Pentose and glucuronate interconversions	53	4	0,044	3,1214	0,0394
	Pyruvic acid; L-Threonine; L-Isoleucine;	Valine, leucine and isoleucine biosynthesis	27	3	0,045	3,1107	0,0350
	Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid;	Ascorbate and aldarate metabolism	45	5	0,045	3,0926	0,1383
	L-Glutamic acid; Pyruvic acid; Butyric acid; Oxoglutaric acid;	Butanoate metabolism	40	4	0,046	3,0843	0,0852
	D-Glucose; Glyceric acid; Pyruvic acid;	Pentose phosphate pathway	32	3	0,046	3,0769	0,0218
	N-Acetyl-L-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; L-Glutamine; L-Glutamic acid; 2-Keto-glutaramic acid	Alanine, aspartate and glutamate metabolism	24	7	0,048	3,0446	0,4122

[a] Total cmpd is the total number of compounds in the pathway.
[b] Hits is the actual matched number from the uploaded data.
[c] Raw p is the original *p*-value calculated from the pathway analysis.
[d] Impact is the pathway impact value calculated from pathway topology analysis.

though deep learning methods are of particular interest in many fields, they necessitate a very large number of patients to be efficiently trained and may therefore not be suitable for small metabolomics datasets obtained on real life patients, such as the one we have used. While obtaining imaging or clinical data concerning several thousands of patients seems achievable, obtaining metabolomics data for that many patients is currently much more complicated. Furthermore, even though some efforts are being made to tackle this issue [60], it is currently impossible to understand which features are responsible for the outcome when using deep-learning clustering techniques. It would therefore be impossible to understand the metabolic differences underlying different patient clusters if deep learning clustering was used.

These considerations raise important questions: in the future, on what basis should decisions be made? On results from a single method? Or on results provided by several methods? In view of the

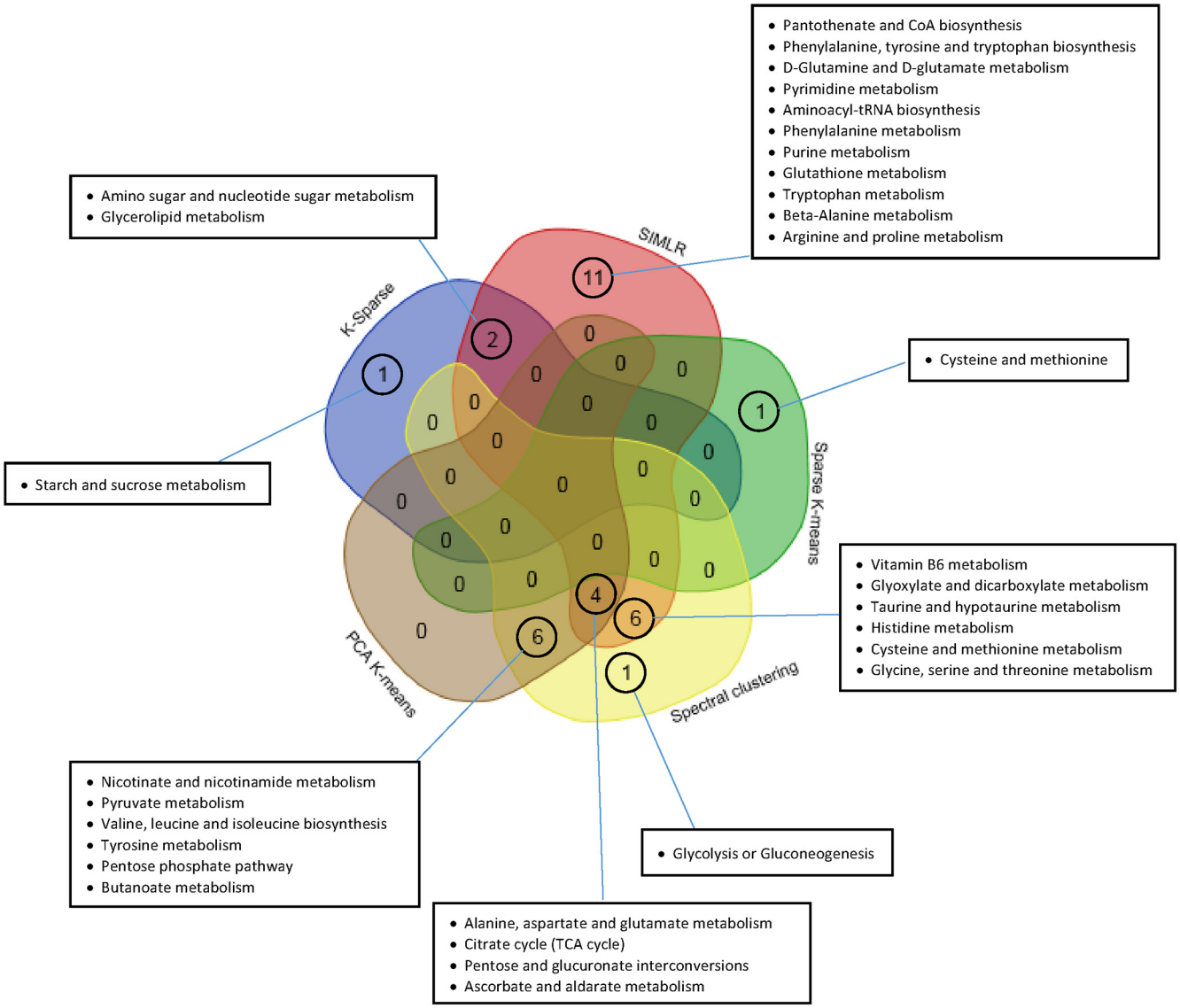

Fig. 4. Venn diagram of pathways that were in common or unique to the five clustering methods.

findings we have highlighted, it seems that decisions should be taken collegially, i.e. based on the results of a set of methods, as at multidisciplinary consultation meetings involving health professionals from different disciplines and whose skills are essential to take decisions ensuring patients the best possible care according to the state of the science.

4.2. From a clinical perspective

From a clinical point of view, the methods were able to highlight three distinct groups of patients with different clinical profiles. Patients identified in cluster 1 may be considered to have the best prognosis, patients in cluster 2 an intermediate prognosis, while patients in cluster 3 may be considered to have the worst prognosis. The results in Table 2 show that the tumors of patients in cluster 1 were predominantly non-invasive and non-proliferative, whereas the tumors of cluster 3 patients were mainly invasive and proliferative. Tumors in cluster 2 were rather invasive but not proliferative, hence the intermediate prognosis. We hypothesize that these patients would have an intermediate (atypical) biological profile, which is why the methods are discordant. We further evidence heterogeneity within the triple-negative BC subpopulation with most of the patients classified in cluster 3. However, a third of the triple-negative patients were in cluster 1 Recent molecular profiling studies of triple-negative BC using parallel sequencing and other "omics" technologies have also uncovered an unexpectedly high level of heterogeneity as well as a number of common features [61,62].

In addition, no significant difference between clusters could be demonstrated in terms of age, histologic type, lymph node involvement, metastasis or survival (OS, SS or RFS). Indeed, with a median follow-up of only 48.5 months, this duration is insufficient to demonstrate a significant difference in terms of OS, SS, or RFS. Nevertheless, it is quite easy to predict that patients in cluster 3 have the highest risk of progression and that, conversely, patients in cluster 1 have the lowest risk of progression. To confirm this intuition and try to reduce this short follow-up limitation, we analyzed simulated survival data obtained with the PREDICT tool. With a 5-year pOS rate at around 75% for cluster 1, 70% for cluster 2 and 60% for cluster 3, *in-silico* analyses have demonstrated their high potential value [28,63,64] and confirmed that patients in cluster 3 have a poorer prognosis [65,66]. One limitation of our study could be the

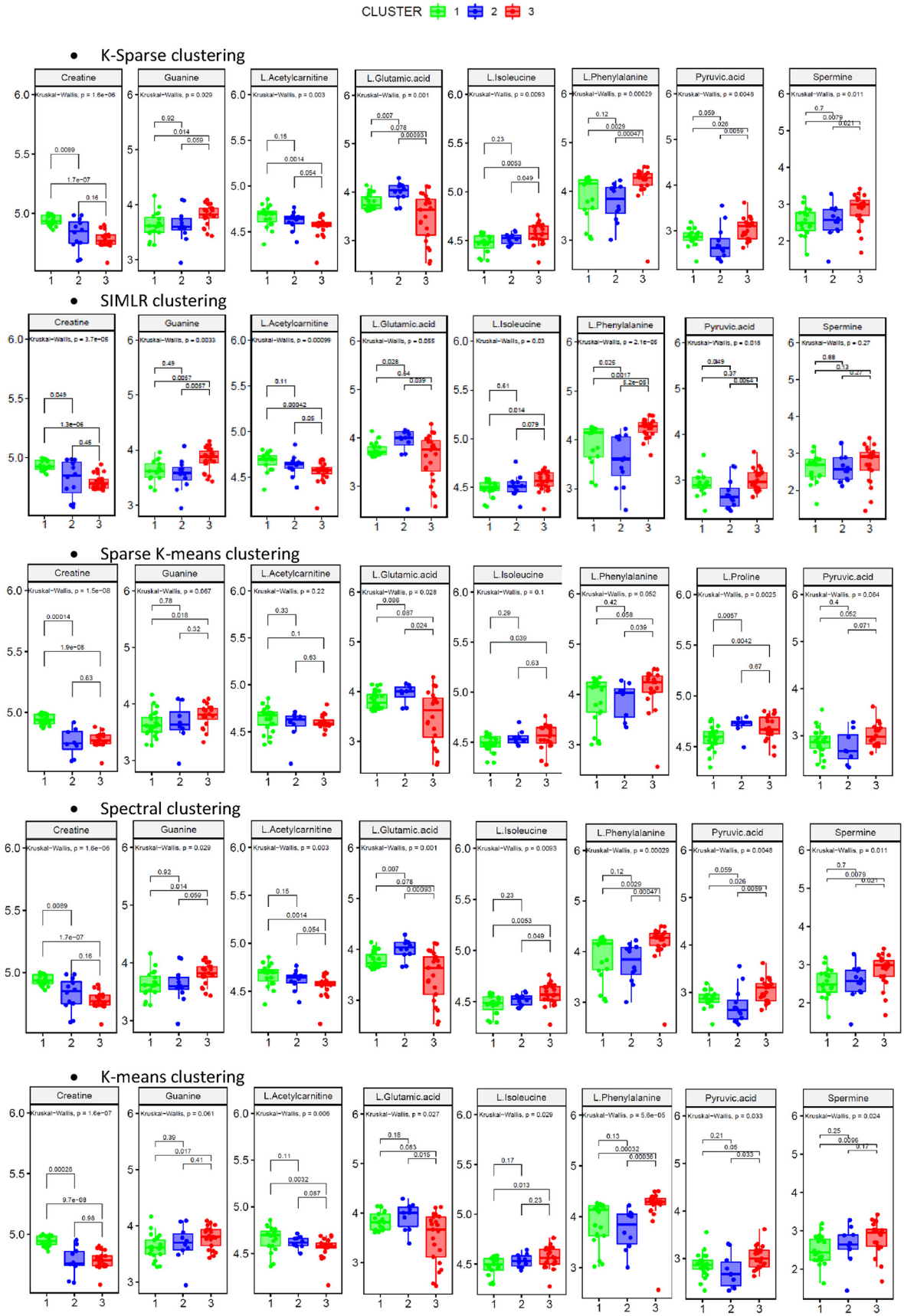

Fig. 5. Boxplot of the 8 metabolites extracted from 5 ML methods.

representativity of our population, e.g. it is recognized that BCs in younger patients (<40 years) are more aggressive [67]. Our study did not include a large number of young patients, which could explain why no significant difference was demonstrated in terms of age between clusters. Similarly, with only three patients with invasive lobular carcinoma (6%), our results did not identify a metabolic signature associated with this phenotype. Previous studies have shown a survival benefit in favor of invasive lobular carcinoma [68,69] and metabolomic studies focused on this particular type of BC could provide valuable biological information. Furthermore, due to the over-representation of hormonal-receptor negative tumors (48%) in our population compared to the literature [70], our population could have had unfavorable prognosis. This bias may result from our method of tumor selection. We decided to analyze frozen samples available in our biobank. Obviously, hormonal-receptor negative, triple-negative, Her-2-positive tumors are more often frozen and stored for further molecular testing and inclusion in clinical trials. In the present study, it is interesting to note that the five methods classified 73% of the patients in the same cluster. Among the 27% of patients classified differently by at least one of the methods, 9.5% of patients were classified heterogeneously by the five methods. Indeed, for each of these 5 patients, three methods classified them in one cluster and 2 others in another cluster without any connection between the types of methods used. Moreover, it is interesting to note that the different methods classified patients, on the one hand, in either the good prognostic cluster or the intermediate prognostic cluster or, on the other, in either the intermediate prognostic cluster or the poor prognostic cluster, but never in the good prognostic cluster or the poor prognostic cluster. A clinical analysis of these 5 patients showed that they had atypical clinical profiles, probably due to particular biological profiles. These atypical profiles would explain why no classification consensus could be highlighted. Overall, ML methods must remain a decision-making tool for the clinician, especially in cases where patients have particular clinical and biological characteristics. To avoid possible medical errors, the final responsibility for the decision lies with the clinician [71].

Finally, the initial clinical objective of this study was to define a metabolomic signature to refine the current classification and help the clinician in his chemotherapy prescription. This paper is the result of methodological research analyzing the best ML methods to develop this new tool. The patients selected were therefore patients eligible for adjuvant chemotherapy. An analysis of the metastatic population could help define a specific signature of metastatic status and/or a signature associated to survival. However, the use of biopsy faces two practical difficulties: 1) the intratumoral and inter-site heterogeneity that could be overcome through the analysis of blood or urine samples; and 2) the amount of material available once the pathologic analyses essential for patient management have been performed. Metabolomic analysis on paraffin slides could facilitate access to specimens and limit the amount of material required.

4.3. From a biological perspective

From a physiological point-of-view, this study extends the molecular stratification of BC to metabolomic profiles. Indeed, our results suggest that dysregulation of metabolic pathways exists between BC subtypes and that a particular amino acid profile characterizes the different BC histologic subtypes. Dysregulations of amino acid metabolism are well-known key events during cancer development [72] and are emerging hallmarks of cancers [73,74]. Amino acids serve not only as building blocks in protein synthesis but also as energy sources favoring cancer cell proliferation and growth [75]. Of interest, we identified significant differences between the BC subtypes of three metabolic pathways (i.e. Glycolysis and lactate production, Glutaminolysis, and amino acid) that play a pivotal role in BC growth [76,77]. Using the five methods, we consistently found that patients in cluster 3 showed higher levels of Guanine, L-Isoleucine, L Methionine, L-Phenylalanine, Pyruvic acid, Spermine and low levels of Creatine, L-Acetylcarnitine and L-Glutamic acid. Our results suggested that these metabolites could be candidate biomarker predictors of poorer prognosis [78–82]. All these results are consistent with the literature [57,83–86].

Given the exploratory nature of our study, we decided to use an FDR rate of 0.25 as a threshold in order to identify relevant candidate pathways (https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/FAQ).

A validation of these pathways, during a study whose main objective will be to evaluate the usefulness of our metabolomics signatures for decision-making, will need to be established with the use of a lower False Discovery Rate or Family Wise Error Rate (<0.05).

Indeed, to meet the biosynthetic needs associated with rapid proliferation, cancer cells must increase the import of nutrients. Two main metabolites are essential for biosynthesis and survival in mammalian cells, and particularly in cancer cells: glucose [87] and glutamine [88]. The increased glucose uptake in tumors compared to other healthy and non-proliferative tissues was first described more than 90 years ago by Otto Warburg [89]. Glucose is the primary energy source of all cells because of its involvement in many processes such as glycolysis or the Krebs cycle [90] in mitochondria. Unlike healthy cells that adapt to available substrates (glucose/fatty acids/proteins), some tumor cells are addicted to glucose. The other important point is that, once metabolized, tumor cells will prefer lactic fermentation to the Krebs cycle.

Lastly, the precise etiology of BC is still unknown even though some genetic, epigenetic and environmental factors have been identified [91]. It has been conclusively demonstrated that cancer cell metabolism is heavily influenced by microenvironmental factors, including nutrient availability. Sullivan and coworkers [92] found that diet affects local nutrient availability. This effect can lead to substantial changes in the metabolism of tumor cells, thereby modifying the response of these cells to drugs targeting metabolism. Drugs capable of inhibiting tumor proliferation may then become ineffective. Therefore, knowledge of microenvironmental nutrient levels is essential to a better understanding of tumor metabolism.

Outcomes for cancer patients vary greatly. The classification of BC into subtypes has been was defined in the literature on the basis of molecular characterization of proteomics (single omic). This has helped improve prognosis and personalized treatment. These considerations have motivated efforts to produce large amounts of multi-omic data such as TCGA [93] and ICGC [94]. However, current algorithms still face challenges and need to integrate omic data [95–98]. Defining BC subtypes using multi-omic data could help to better understand some of the dark areas that still persist in the field of tumor mechanisms in order to offer even more personalized treatments.

5. Conclusion

In the era of personalized medicine, OMICS science (genomics, transcriptomics, proteomics, and metabolomics) must contribute to the quest for cancer-specific biomarkers. The present study argues in favor of further research in this domain. Metabolomics is emerging as a relevant and promising tool for the classification of BC to enable more precise diagnosis [54,99–101]. Even though it is less accurate than the targeted approach, untargeted metabolomics nevertheless permits identification and quantification of a

vast number of major metabolites. Thus, this approach presents a particular interest in the search for new candidate biomarkers [102–104] and could be applied in everyday medical practice given that the cost and duration of metabolomic analyses are relatively low. However, due to the retrospective design of our study and the small number of patients recruited, our results need to be validated in a larger cohort and in the context of a prospective clinical trial.

Funding

The authors declare no competing financial interests.

CrediT authorship contribution statement

Jocelyn Gal: Methodology, Formal analysis, Writing - original draft. **Caroline Bailleux:** Writing - original draft. **David Chardin:** Software, Writing - original draft. **Thierry Pourcher:** Conceptualization, Writing - review & editing. **Julia Gilhodes:** . **Lun Jing:** . **Jean-Marie Guigonis:** Methodology, Writing - review & editing. **Jean-Marc Ferrero:** Data curation. **Gerard Milano:** Writing - review & editing. **Baharia Mograbi:** Writing - review & editing. **Patrick Brest:** Writing - review & editing. **Yann Chateau:** . **Olivier Humbert:** Conceptualization, Writing - review & editing. **Emmanuel Chamorey:** Supervision, Methodology, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge support from the Centre Antoine Lacassagne, TIRO Unit, University Côte d'Azur and the Departmental Council of the Alpes Maritimes, France.

The authors sincerely thank Mrs. Clair Della Vedova for her help in developing the figures.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.05.021.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. CA Cancer J Clin 2017;67:7–30.
[2] Perou CM, Jeffrey SS, van de Rijn M, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA 1999;96:9212–7.
[3] Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature 2000;405:827–36.
[4] Pandey A, Mann M. Proteomics to study genes and genomes. Nature 2000;405:837–46.
[5] Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–52.
[6] Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001;98:10869–74.
[7] Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 2003;100:8418–23.
[8] Witten DM, Tibshirani R. A framework for feature selection in clustering. J Am Stat Assoc 2010;105:713–26.
[9] Candido Dos Reis FJ, Wishart GC, Dicks EM, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. Breast Cancer Res 2017;19:58.
[10] Wishart GC, Azzato EM, Greenberg DC, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res 2010;12:R1.
[11] Ross JS. Multigene predictors in early-stage breast cancer: moving in or moving out?. Expert Rev Mol Diagn 2008;8:129–35.
[12] Ross JS, Hatzis C, Symmans WF, et al. Commercialized multigene predictors of clinical outcome for breast cancer. Oncologist 2008;13:477–93.
[13] Buyse M, Loi S, van't Veer L,, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 2006;98:1183–92.
[14] Cao Y, DePinho RA, Ernst M, Vousden K. Cancer research: past, present and future. Nat Rev Cancer 2011;11:749–54.
[15] Ehmann F, Caneva L, Prasad K, et al. Pharmacogenomic information in drug labels: European Medicines Agency perspective. Pharmacogenomics J 2015;15:201–10.
[16] McShane LM, Polley MY. Development of omics-based clinical tests for prognosis and therapy selection: the challenge of achieving statistical robustness and clinical utility. Clin Trials 2013;10:653–65.
[17] van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.
[18] Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365:671–9.
[19] Wesolowski R, Ramaswamy B. Gene expression profiling: changing face of breast cancer classification and management. Gene Expr 2011;15:105–15.
[20] Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer?. Nat Rev Cancer 2012;12:323–34.
[21] Hsu PP, Sabatini DM. Cancer cell metabolism: Warburg and beyond. Cell 2008;134:703–7.
[22] McClellan J, King MC. Genetic heterogeneity in human disease. Cell 2010;141:210–7.
[23] Cannon WB. The wisdom of the body. 2nd ed. Oxford, England: Norton & Co.; 1939.
[24] Roberts LD, Souza AL, Gerszten RE, Clish CB. Targeted metabolomics. Curr Protoc Mol Biol 2012. Chapter 30: Unit 30 32 31-24.
[25] Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted metabolomics strategies-challenges and emerging directions. J Am Soc Mass Spectrom 2016;27:1897–905.
[26] Vinayavekhin N, Saghatelian A. Untargeted metabolomics. Curr Protoc Mol Biol 2010. Chapter 30: Unit 30 31 31-24.
[27] Camacho DM, Collins KM, Powers RK, et al. Next-generation machine learning for biological networks. Cell 2018;173:1581–92.
[28] Gal J, Milano G, Ferrero JM, et al. Optimizing drug development in oncology by clinical trial simulation: why and how? Brief Bioinform 2017.
[29] Yu MK, Ma J, Fisher J, et al. Visible machine learning for biomedicine. Cell 2018;173:1562–5.
[30] Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015;349:255–60.
[31] Tang P, Tse GM. Immunohistochemical surrogates for molecular classification of breast carcinoma: A 2015 update. Arch Pathol Lab Med 2016;140:806–14.
[32] Katajamaa M, Oresic M. Processing methods for differential analysis of LC/MS profile data. BMC Bioinf 2005;6:179.
[33] Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinf 2010;11:395.
[34] Xia J, Mandal R, Sinelnikov IV, et al. MetaboAnalyst 2.0 – a comprehensive server for metabolomic data analysis. Nucleic Acids Res 2012;40:W127–133.
[35] Irizarry RA, Wang C, Zhou Y, Speed TP. Gene set enrichment analysis made simple. Stat Methods Med Res 2009;18:565–75.
[36] Saxena A, Prasad M, Gupta A, et al. A review of clustering techniques and developments. Neurocomputing 2017;267:664–81.
[37] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J Royal Stat Soc: Series B (Statistical Methodol) 2001;63:411–23.
[38] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons; 2009.
[39] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.
[40] Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1979:224–7.
[41] Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat-Theory Methods 1974;3:1–27.
[42] Wang B, Zhu J, Pierson E, et al. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods 2017;14:414–6.
[43] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics; 2007. p. 1027–35.
[44] Lloyd S. Least squares quantization in PCM. IEEE Trans. Inform. Theory 1982;28(2):129–37. https://doi.org/10.1109/TIT.1982.1056489.
[45] Steinhaus H. Sur la division des corps materiels en parties. Bull. Acad. Polon. Sci., C1. III 1956;IV:801–4.
[46] Ng AY, Jordan MI, Weiss Y. Analysis and an algorithm. In: Advances in neural information processing systems. On spectral clustering; 2002. p. 849–56.
[47] Von Luxburg U. A tutorial on spectral clustering. Stat Comput 2007;17:395–416.

[48] Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc: Ser B (Methodol) 1996:267–88.
[49] Gilet C, Deprez M, Caillau J-B, Barlaud M. Clustering with feature selection using alternating minimization, Application to computational biology. arXiv preprint arXiv:1711.02974 2017.
[50] Lvd Maaten, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.
[51] Team RCR. A language and environment for statistical. Computing 2013.
[52] Witten DM, Tibshirani R. sparcl: Perform sparse hierarchical clustering and sparse k-means clustering. R package version 2013;1.
[53] Wishart GC, Bajdik CD, Azzato EM, et al. A population-based validation of the prognostic model PREDICT for early breast cancer. Eur J Surg Oncol 2011;37:411–7.
[54] Beger RD. A review of applications of metabolomics in cancer. Metabolites 2013;3:552–74.
[55] Gunther UL. Metabolomics biomarkers for breast cancer. Pathobiology 2015;82:153–65.
[56] McCartney A, Vignoli A, Biganzoli L, et al. Metabolomics in breast cancer: a decade in review. Cancer Treat Rev 2018;67:88–96.
[57] Silva C, Perestrelo R, Silva P, et al. Breast cancer metabolomics: from analytical platforms to multivariate data analysis. A Review. Metabolites 2019:9.
[58] Asiago VM, Alvarado LZ, Shanaiah N, et al. Early detection of recurrent breast cancer using metabolite profiling. Cancer Res 2010;70:8309–18.
[59] Cardoso MR, Santos JC, Ribeiro ML, et al. A Metabolomic approach to predict breast cancer behavior and chemotherapy response. Int J Mol Sci 2018;19.
[60] Karim MR, Beyan O, Zappa A, et al. Deep learning-based clustering approaches for bioinformatics. Brief Bioinform 2020.
[61] Bianchini G, Balko JM, Mayer IA, et al. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. Nat Rev Clin Oncol 2016;13:674–90.
[62] Mills MN, Yang GQ, Oliver DE, et al. Histologic heterogeneity of triple negative breast cancer: A national cancer centre database analysis. Eur J Cancer 2018;98:48–58.
[63] Belkacemi Y, Hanna NE, Besnard C, et al. Local and regional breast cancer recurrences: salvage therapy options in the new era of molecular subtypes. Front Oncol 2018;8:112.
[64] Buonaguro FM, Caposio P, Tornesello ML, et al. Cancer diagnostic and predictive biomarkers 2018. Biomed Res Int 2019;2019:3879015.
[65] Ponde NF, Zardavas D, Piccart M. Progress in adjuvant systemic therapy for breast cancer. Nat Rev Clin Oncol 2018.
[66] Senkus E, Kyriakides S, Ohno S, et al. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2015;26(Suppl 5):v8–30.
[67] Assi HA, Khoury KE, Dbouk H, et al. Epidemiology and prognosis of breast cancer in young women. J Thorac Dis 2013;5(Suppl 1):S2–8.
[68] Wang K, Zhu GQ, Shi Y, et al. Long-term survival differences between T1–2 invasive lobular breast cancer and corresponding ductal carcinoma after breast-conserving surgery: A propensity-scored matched longitudinal cohort study. Clin Breast Cancer 2019;19:e101–15.
[69] Wasif N, Maggard MA, Ko CY, Giuliano AE. Invasive lobular vs. ductal breast cancer: a stage-matched comparison of outcomes. Ann Surg Oncol 2010;17:1862–9.
[70] Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. World J Clin Oncol 2014;5:412–24.
[71] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.
[72] Pavlova NN, Thompson CB. The emerging hallmarks of cancer metabolism. Cell Metab 2016;23:27–47.
[73] Hainaut P, Plymoth A. Targeting the hallmarks of cancer: towards a rational approach to next-generation cancer therapy. Curr Opin Oncol 2013;25:50–1.
[74] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–74.
[75] Li Z, Zhang H. Reprogramming of glucose, fatty acid and amino acid metabolism for cancer progression. Cell Mol Life Sci 2016;73:377–92.
[76] DeBerardinis RJ, Chandel NS. Fundamentals of cancer metabolism. Sci Adv 2016;2:e1600200.
[77] Haukaas TH, Euceda LR, Giskeodegard GF, Bathen TF. Metabolic portraits of breast cancer by HR MAS MR spectroscopy of intact tissue samples. Metabolites 2017:7.
[78] Jeon H, Kim JH, Lee E, et al. Methionine deprivation suppresses triple-negative breast cancer metastasis in vitro and in vivo. Oncotarget 2016;7:67223–34.
[79] Melone MAB, Valentino A, Margarucci S, et al. The carnitine system and cancer metabolic plasticity. Cell Death Dis 2018;9:228.
[80] Thomas TJ, Thomas T. Cellular and animal model studies on the growth inhibitory effects of polyamine analogues on breast cancer. Med Sci (Basel) 2018:6.
[81] Xiao F, Wang C, Yin H, et al. Leucine deprivation inhibits proliferation and induces apoptosis of human breast cancer cells via fatty acid synthase. Oncotarget 2016;7:63679–89.
[82] Zuo Y, Ulu A, Chang JT, Frost JA. Contributions of the RhoA guanine nucleotide exchange factor Net1 to polyoma middle T antigen-mediated mammary gland tumorigenesis and metastasis. Breast Cancer Res 2018;20:41.
[83] Lecuyer L, Dalle C, Lyan B, et al. Plasma metabolomic signatures associated with long-term breast cancer risk in the SU.VI.MAX prospective cohort. Cancer Epidemiol Biomarkers Prev 2019.
[84] Oikari S, Kettunen T, Tiainen S, et al. UDP-sugar accumulation drives hyaluronan synthesis in breast cancer. Matrix Biol 2018;67:63–74.
[85] Pan H, Xia K, Zhou W, et al. Low serum creatine kinase levels in breast cancer patients: a case-control study. PLoS One 2013;8:e62112.
[86] Phannasil P, Ansari IH, El Azzouny M, et al. Mass spectrometry analysis shows the biosynthetic pathways supported by pyruvate carboxylase in highly invasive breast cancer cells. Biochim Biophys Acta Mol Basis Dis 2017;1863:537–51.
[87] Mason EF, Rathmell JC. Cell metabolism: an essential link between cell growth and apoptosis. Biochim Biophys Acta 2011;1813:645–54.
[88] Hensley CT, Wasti AT, DeBerardinis RJ. Glutamine and cancer: cell biology, physiology, and clinical opportunities. J Clin Invest 2013;123:3678–84.
[89] Warburg O, Wind F, Negelein E. The metabolism of tumors in the body. J Gen Physiol 1927;8:519–30.
[90] Anderson NM, Mucka P, Kern JG, Feng H. The emerging role and targetability of the TCA cycle in cancer metabolism. Protein Cell 2018;9:216–37.
[91] Fernandez MF, Reina-Perez I, Astorga JM, et al. Breast Cancer and Its Relationship with the Microbiota. Int J Environ Res Public Health 2018;15.
[92] Sullivan MR, Danai LV, Lewis CA, et al. Quantification of microenvironmental metabolites in murine cancers reveals determinants of tumor nutrient availability. Elife 2019:8.
[93] Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–8.
[94] Zhang J, Baran J, Cros A, et al. International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. Database (Oxford) 2011;2011:bar026.
[95] Mitra S, Saha S. A multiobjective multi-view cluster ensemble technique: Application in patient subclassification. PLoS One 2019;14:e0216904.
[96] Ramazzotti D, Lal A, Wang B, et al. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nat Commun 2018;9:4453.
[97] Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res 2018;46:10546–62.
[98] Wu C, Zhou F, Ren J, et al. A selective review of multi-level omics data integration using variable selection. High Throughput 2019:8.
[99] Armitage EG, Barbas C. Metabolomics in cancer biomarker discovery: current trends and future perspectives. J Pharm Biomed Anal 2014;87:1–11.
[100] Bennett DA, Waters MD. Applying biomarker research. Environ Health Perspect 2000;108:907–10.
[101] Vermeersch KA, Styczynski MP. Applications of metabolomics in cancer research. J Carcinog 2013;12:9.
[102] Jacob M, Lopata AL, Dasouki M, Abdel Rahman AM. Metabolomics toward personalized medicine. Mass Spectrom Rev 2017.
[103] Trivedi DK, Hollywood KA, Goodacre R. Metabolomics for the masses: The future of metabolomics in a personalized world. New Horiz Transl Med 2017;3:294–305.
[104] Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nat Rev Drug Discov 2016;15:473–84.

ANNEXE 2

Médecine Nucléaire 45 (2021) 4–12

Disponible en ligne sur ScienceDirect www.sciencedirect.com

Elsevier Masson France EM|consulte www.em-consulte.com

Article original

Métabolomique et imagerie TEP-FDG des cancers du sein

Metabolomics and FDG-PET-CT in breast cancer

D. Chardin [a,*], T. Pourcher [b], J. Gal [c], C. Bailleux [d], J.M. Guigonis [b], J. Darcourt [a], L. Arnould [e], O. Humbert [a]

[a] *Service de médecine nucléaire, centre Antoine-Lacassagne, 33, avenue de Valombrose, 06189 Nice, France*
[b] *UMRE 4320, laboratoire TIRO, CEA, université de Nice, 06189 Nice, France*
[c] *Unité de biostatistique, centre Antoine-Lacassagne, 33, avenue de Valombrose, 06189 Nice, France*
[d] *Service oncologie, centre Antoine-Lacassagne, 33, avenue de Valombrose, 06189 Nice, France*
[e] *Département de biologie et pathologie des tumeurs, centre Georges-François Leclerc, 1, rue du Professeur Marion, 21000 Dijon, France*

INFO ARTICLE

Historique de l'article :
Reçu le 24 septembre 2019
Accepté le 30 mars 2020
Disponible sur Internet le 24 mai 2020

Mots clés :
FDG
Métabolomique
Cancer du sein

RÉSUMÉ

Le métabolisme tumoral est étroitement lié à la tumorogenèse et varie selon le phénotype tumoral. Les techniques de métabolomique peuvent fournir des données quantitatives concernant un grand nombre de métabolites et permettent l'analyse de nombreuses voies métaboliques intriquées. La tomographie par émission (TEP) de positons au ^{18}F-fluorodéoxyglucose (FDG) permet une mesure in vivo de l'avidité cellulaire en glucose. L'objectif de ce travail était de rechercher et comprendre les corrélations existantes entre données de métabolomique et intensité de captation du FDG dans le cadre du cancer du sein. Soixante-dix échantillons tumoraux provenant de patientes présentant un cancer du sein et pour lesquelles les résultats d'une TEP-FDG préthérapeutique étaient disponibles ont été analysés selon un protocole de métabolomique, non ciblée par chromatographie liquide et couplée à une spectrométrie de masse. Les échantillons ont été séparés en deux groupes selon l'intensité de fixation tumorale du FDG en prenant la valeur médiane de la SUV_{max} pour dichotomiser. Huit cent cinquante-quatre métabolites ont été identifiés. Une analyse discriminante des moindres carrés partiels et une analyse de classification supervisée ont permis de créer un modèle permettant de prédire l'avidité en FDG à partir des données de métabolomique avec une précision de 0,73 à 0,77. Les métabolites corrélés à l'intensité de fixation du FDG étaient, entre autres, la glutathione, des acides aminés tels que le glutamate, la proline ou la tyrosine, l'acetyl-carnitine, des métabolites de la voie de la kynurenine et des polyamines telles que la N1,N12-diacetylspermine. Ces métabolites ont déjà été rapportés comme marqueurs de l'agressivité tumorale. Une corrélation directe entre avidité en FDG et glycolyse n'a pas pu être mise en évidence, cependant de nombreux signes indirects montraient une plus forte activité glycolytique dans les tumeurs avides en FDG. L'analyse de métabolites inconnus, mise en évidence par cette approche, pourrait permettre l'identification de nouveaux biomarqueurs d'intérêt.

© 2020 Elsevier Masson SAS. Tous droits réservés.

Keywords:
FDG
Metabolomics
Breast cancer

ABSTRACT

Cancer metabolism is an essential aspect of tumorogenesis, as cancer cells have increased energy requirements in comparison to normal cells. Metabolomic techniques can provide quantitative data for a large number of small molecules in tissues and enable the analysis of multiple intricate metabolic pathways. Positron emission tomography (PET) using ^{18}F-Fluorodeoxyglucose (FDG) enables in vivo analysis of glycolysis and is widely used in oncology. High tumor FDG uptake is a prognostic factor in breast cancer and has been associated with tumor aggressively. Seventy breast cancer samples obtained from untreated patients who had underwent FDG-PET imagery were analyzed through an untargeted metabolomic approach using liquid chromatography-mass spectroscopy (LC-MS) to study possible correlations between metabolomic data and FDG uptake. Tumors were split into two groups depending

* Auteur correspondant.
Adresse e-mail : chardindj@gmail.com (D. Chardin).

https://doi.org/10.1016/j.mednuc.2020.03.002
0928-1258/© 2020 Elsevier Masson SAS. Tous droits réservés.

on avidity for FDG as measured with PET. The Compound Discoverer 4.0 software enabled identification of 854 metabolites. PLSDA based models predicted FDG uptake with an accuracy ranging from 0,73 to 0,77. Selected metabolites varied depending on the use of scaling or log transformation. Metabolites correlated with tumor FDG uptake were, among others, glutathione, amino-acids such as glutamate, proline or tyrosine, L-acetyl-carnitine, metabolites from the kynurenine pathway such as L-kynurenine or formyl-kynurenine and polyamines such as N1,N12-diacetylspermine or N1-acetylspermine. These metabolites have been previously shown to reflect cancer aggressivity. The correlation between the glycolytic pathway activation and tumor FDG uptake could not be directly assessed but indirect signs showed a higher glycolytic activity in tumours presenting a higher FDG uptake. Studying new metabolites identified through this process could enable a better understanding of tumor metabolism and identification of new biomarkers.

© 2020 Elsevier Masson SAS. All rights reserved.

1. Introduction

De nombreuses études ont montré que le métabolisme tumoral est différent du métabolisme des tissus sains. La reprogrammation du métabolisme dans le cadre du cancer est un processus complexe, non encore complètement élucidé, impliquant des mutations, des modifications épigénétiques et l'influence du micro-environnement [1].

L'une des différences les plus connues entre le métabolisme tumoral et le métabolisme des tissus sains a été mise en évidence par O. Warburg dans les années 1920 et est nommée « l'effet Warburg » [2]. Cet effet correspond au fait que les cellules tumorales utilisent préférentiellement la glycolyse suivie de la fermentation lactique plutôt que la phosphorylation oxydative comme voie énergétique et ce indépendamment de la disponibilité en oxygène ou des autres substrats énergétiques. Ce phénomène est appelé « Effet Warburg » ou « glycolyse aérobie » [3].

Les mécanismes à l'origine de l'effet Warburg ne sont pas encore complètement élucidés. Plusieurs hypothèses ont été proposées pour expliquer les bénéfices potentiels de l'hyperactivation de la glycolyse aérobie et la formation de lactates pour les cellules tumorales malgré le faible rendement énergétique de cette voie métabolique, mais aucune n'a été confirmée [4].

Ce faible rendement énergétique implique une forte consommation de glucose. Pour faciliter le passage intracellulaire du glucose, les cellules tumorales expriment un grand nombre de transporteurs membranaires du glucose (GLUT) et ce même en situation de jeûne et d'insulinémie faible. Cette caractéristique est ce qui a permis l'essor de la tomographie par émission de positons (TEP) TEP-TDM au ^{18}F-deoxygluxose-6-phosphate (FDG) en cancérologie [5], notamment dans le cadre du cancer du sein [6]. Celle-ci permet de mesurer in vivo l'absorption cellulaire de glucose [7].

Compte tenu des connaissances concernant l'effet Warburg, il est globalement admis que la captation tumorale du FDG est corrélée à l'activité glycolytique tumorale, bien que ceci n'ai jamais été démontré expérimentalement. Par ailleurs, l'intensité de captation tumorale du FDG est un facteur pronostique connu pour de nombreux cancers, notamment le cancer du sein [8].

La captation tumorale du FDG est hétérogène. Elle varie selon le type tumoral et entre différentes lésions métastatiques chez un même patient. Ceci est particulièrement vrai dans le cadre du cancer du sein [9]. D'autres voies métaboliques que la glycolyse aérobie sont vraisemblablement activées au sein du tissus tumoral et on peut supposer que les tumeurs présentant une faible avidité pour le FDG utilisent d'autres substrats énergétiques tels que les acides gras ou la glutamine plutôt que le glucose.

La métabolomique est une nouvelle approche permettant de quantifier un grand nombre de métabolites dans un fluide ou un tissu par le biais d'analyses en spectrométrie de masse [10]. Elle permet ainsi d'étudier l'état d'activation d'un grand nombre de voies métaboliques par le biais d'analyses statistiques et par confrontation à des bases de données. Les données de métabolomique ont l'avantage de refléter la dernière étape des différentes cascades de signaux exercés sur les tissus. Cette approche a notamment déjà apporté des résultats pertinents dans le cadre du cancer du sein [11].

L'objectif de ce travail a été de rechercher des corrélations entre des données de métabolomique et l'intensité de captation tumorale du FDG puis analyser les métabolites corrélés afin d'identifier des facteurs pronostiques et des radiopharmaceutiques potentiellement utiles en cas de faible captation du FDG.

2. Matériel et méthodes

2.1. Population de l'étude

Cette étude a porté sur des échantillons tumoraux prélevés chez 70 patientes atteintes de cancers du sein au stade localisé avec ou sans extension ganglionnaire locorégionale, présentant une indication de traitement néoadjuvant. Ces patientes ont été incluses de manière prospective au centre de cancérologie de Dijon entre novembre 2006 et novembre 2013. Pour chaque patiente, le poids et la taille ont été recueillis, ainsi que les antécédents familiaux néoplasiques, les antécédents obstétricaux et le statut ménopausique. Le stade tumoral, le grade histologique et le phénotype tumoral ont également été recueillis.

2.2. Mesure de la captation tumorale du ^{18}F-deoxygluxose-6-phosphate

Une TEP-TDM au ^{18}FDG a été réalisée avant tout traitement pour les 70 patientes incluses, en moyenne 14 (±10) jours après la biopsie.

Deux systèmes TEP-TDM ont été utilisés : un système Gemini GXL entre novembre 2006 et décembre 2010 et un système Gemini TF PET-CT entre décembre 2010 et décembre 2013 (Philips Medical Systems, Eindhoven, Pays-Bas). Les patientes avaient comme instruction de jeûner au moins 6 h avant l'injection intraveineuse de 5 MBq/kg de FDG pour les examens réalisés sur le système Gemini GXL et de 3 MBq/kg pour les examens réalisés sur le système Gemini TF. L'acquisition TEP-TDM a été réalisée 60 min après injection. Les images TEP ont été corrigées de la décroissance radioactive, de la diffusion, des coïncidences fortuites et de

l'atténuation avant la reconstruction par algorithme itératif : 3D-RAMLA (GEMINI GXL) ou 3D OSEM (GEMINI TF).

La captation tumorale a été quantifiée relativement par la mesure de la SUV_{max} (*Maximal Standard Uptake Value*) selon la formule suivante :

$$SUV_{max} = \frac{\text{Activité dans le voxel le plus intense de la tumeur kBq.ml}^{-1} \times \} \text{ Masse totale g}}{\text{Activité injectée kBq}}$$

Les patientes ont ensuite été séparées en deux groupes selon les valeurs de SUV mesurées : un groupe « tumeur peu avide en FDG » correspondant aux tumeurs présentant une valeur de SUV_{max} inférieure à la médiane et un groupe « tumeur avide en FDG » correspondant aux tumeurs présentant une valeur de SUV_{max} supérieure à la médiane.

2.3. Biopsies

Lors de la biopsie préthérapeutique sous contrôle échographique, 1 à 3 carottes complémentaires de tissu ont été prélevées. Les échantillons tumoraux ont été congelés et conservés à −80 °C jusqu'aux analyses.

2.4. Préparation des échantillons

Afin d'extraire les métabolites du tissu tumoral, chaque échantillon a été plongé encore congelé dans un tube Eppendorf contenant 1 mL de méthanol puis broyé manuellement avec un piston. Les tubes ont ensuite été placés dans un congélateur à −20 °C durant au moins 12 h. Le lendemain, les tubes ont été centrifugés à 13 000 rpm pendant 15 min puis les surnageants ont été extraits et placés dans un nouveau tube. Le méthanol a été évaporé en utilisant un speed-Vac pendant 10 h. Une fois séchés, les échantillons ont été conservés dans un congélateur à −20 °C jusqu'à l'analyse de spectrométrie de masse couplée à la chromatographie liquide (CL-SM). Avant cette analyse, les métabolites ont été re-suspendus dans 100 µL d'un mélange 50 % acétonitrile/50 % eau, compatible avec l'analyse CL-SM.

2.5. Analyses de spectrométrie de masse couplée à la chromatographie liquide (CL-SM)

La quantification des différents métabolites a été réalisée en spectrométrie de masse couplée à la chromatographie en phase liquide.

L'étape de chromatographie liquide a été réalisée avec un système DIONEX Ultimate 3000 HPLC (Thermo Fisher Scientific). L'analyse de spectrométrie de masse était réalisée sur un système Q Exactive Plus Orbitrap (Thermo Scientific) avec une source d'ionisation de type électronébuliseur chauffé (HESI II), en mode positif (tension de pulvérisation à 3800 V) et négatif (tension de pulvérisation à 2500 V). Les paramètres détaillés sont décrits en annexe.

2.6. Traitement informatique des données de spectrométrie de masse couplée à la chromatographie liquide

Les données de spectrométrie de masse couplée à la chromatographie liquide ont été traitées avec le logiciel Compound Discoverer (Version 3.0, ThermoFischer). Les données issues des modes d'ionisation positifs et négatifs ont été traitées séparément. Les paramètres utilisés sont décrits en annexe.

Les résultats obtenus en mode positif et négatif ont été combinés. Quand un métabolite était identifié dans les deux modes, le mode permettant d'obtenir l'intensité moyenne la plus élevée a été privilégié.

Les pics correspondant à des contaminants d'origine pharmacologique, tels que la lidocaïne, ont été supprimés avant les analyses statistiques.

2.7. Préparation des données

Les intensités de pics ont été normalisées en fonction de la somme de l'intensité des pics pour chaque échantillon. Ce mode de normalisation suppose que le courant ionique total est équivalent pour chaque échantillon.

Les données ont été centrées et réduites selon deux méthodes avant les analyses par classification : une réduction classique, par rapport à l'écart type de chaque variable, une réduction selon Pareto, par rapport à la racine carrée de l'écart type de chaque variable. La réduction classique permet d'égaliser l'importance potentielle de chaque métabolite. Elle présente l'inconvénient d'être plus sensible aux erreurs de mesure. La méthode de Pareto limite l'impact des métabolites de grande intensité sur la classification tout en préservant la structure des données [12].

2.8. Analyses statistiques

Les analyses statistiques ont été réalisées avec le logiciel MetaboAnalyst (Version 4.0, https://www.metaboanalyst.ca/).

2.9. Analyses par classification non supervisée

Une analyse en composantes principales (ACP) a été réalisée. Cette analyse permet de rechercher des groupes d'échantillons présentant des caractéristiques similaires sans prendre en compte d'a priori concernant la classification. On peut ainsi avoir une vue globale des résultats et identifier d'éventuels échantillons à caractéristiques aberrantes.

2.10. Analyses par classification supervisée

Une analyse discriminante des moindres carrés partiels a été réalisée pour évaluer si les données de métabolomique peuvent prédire l'avidité en FDG des tumeurs. Cette méthode, fréquemment utilisée en métabolomique, recherche des métabolites permettant de regrouper les échantillons selon les groupes définis par l'utilisateur. Elle attribue pour cela un poids à chaque métabolite et crée un modèle prédictif. Ce modèle prédictif peut être étoffé en augmentant le nombre de composantes explicatives. Cette méthode présente l'avantage d'être simple et validée.

Les métabolites présentant un poids important dans les modèles prédictifs établis ont fait l'objet d'analyses univariées.

2.11. Analyses univariées

Pour chaque métabolite identifié, un test *t* a été réalisé pour comparer les moyennes d'intensité dans le groupe « peu avide en FDG » et le groupe « avide en FDG ». Des analyses par courbe sensibilité/spécificité ont également été réalisées.

Les métabolites présentant des différences statistiquement significatives ont été testés dans une seconde base de données de métabolomique portant sur 30 patientes prises en charge pour des tumeurs du sein pour lesquelles des données de TEP-TDM au ^{18}FDG réalisées avant tout traitement étaient disponibles.

2.12. Validations de l'identification des métabolites d'intérêts

Les métabolites d'intérêts issus des analyses statistiques ont été validés en termes de spectres de masses de premier ordre (MS1) et de second ordre (MS2) par comparaison des spectres obtenus avec ceux des bases de données METLIN. Le MS1 correspond au spectre

de masse issu du premier analyseur du spectromètre de masse en tandem, il contient les pics correspondants au métabolite d'intérêt sous différentes formes d'ionisation avec son profil isotopique. Le MS2 correspond au spectre de masse issu du second analyseur et contient les pics correspondant aux fragments obtenus après passage du métabolite dans la chambre de collision. Si plusieurs métabolites peuvent être isobares, leurs fragments respectifs sont susceptibles de présenter des masses différentes, ce qui permet de les différencier.

3. Résultats

3.1. Caractéristiques des patientes incluses

Les caractéristiques des patientes incluses sont résumées dans le Tableau 1. La plupart des patientes présentaient des tumeurs de stade II (taille tumorale entre 2 et 5 cm ou atteinte ganglionnaire régionale).

3.2. Mesure de la captation tumorale du Fluorodeoxyglucose

Les valeurs de SUV_{max} des tumeurs ont pu être mesurées à partir de TEP-TDM au ^{18}FDG pour les 70 patientes. Un exemple de segmentation est présenté dans la Fig. 1. La valeur médiane de la SUV_{max} était de 6,78 (minimum 1,7 — maximum 50,5).

Tableau 1
Principales caractéristiques des patientes incluses. L'âge, le poids, la taille et l'indice de masse corporelle sont exprimés sous forme de moyennes et des extrêmes ou en valeurs absolues.
Patient characteristics. Values are given as mean and ranges or absolute values.

Caractéristique	
Âge	48 (26–73)
Poids	65,6 (53–91)
Taille	161 (152–170)
Indice de masse corporelle	24,8 (18,8–41,4)
Au moins un antécédent de grossesse	59/70
Ménopause	30/70
Stade	
I	1
II	65
III	4
Stade T (taille tumorale)	
1	(<2 cm) : 3
2	(2–5 cm) : 60
3	(>5 cm) : 7
Stade N (atteinte ganglionnaire)	
0	21
1	48
2	1
Grade SBR	
1	6
2	32 (inconnu : 3)
3	29
Détail SBR : mitoses	
1	26
2	22 (inconnu : 5)
3	17
Détail SBR : architecture	
1	0
2	16 (inconnu : 5)
3	49
Détail SBR : atypies	
1	3
2	24 (inconnu : 5)
3	38
Phénotype	
Luminal	32
HER2	19
Triple négatif	19

La SUV_{max} n'était pas statistiquement associée à l'âge, à l'indice de masse corporelle, aux antécédents de grossesse, au statut ménopausique, à la taille tumorale ($p = 0,169$) ou au stade N.

Une SUV_{max} tumorale élevée était significativement associée à un grade histologique de Scarff Bloom Richardson élevé (SBR) ($p = 0,0017$), au nombre de mitoses ($p = 0,0036$), au nombre d'atypies cellulaires ($p = 0,016$) et au phénotype tumoral ($p = 6E-05$ après analyse de variance).

La SUV_{max} était plus faible pour les phénotypes « luminaux » (moyenne = 6,3), intermédiaire (moyenne = 7,78) pour les phénotypes « HER2 positifs » et plus élevée pour les phénotypes « triples négatifs » (moyenne = 16,24), ce qui concorde avec les résultats déjà rapportés dans la littérature [9].

3.3. Métabolites identifiés

L'utilisation du logiciel « Compound Discoverer » après analyse de CL-SM a permis d'individualiser 3940 pics dont 345 correspondaient à des métabolites identifiés non redondants en mode négatif et 7775 pics dont 625 correspondaient à des métabolites identifiés non redondants en mode positif. La combinaison des deux bases a permis de conserver 854 pics correspondant à des métabolites uniques.

Les pics correspondant à la 3-hydroxylidocaine, au bisoprolol, au bromazépam, au citalopram, à la codéine, à l'hydrochlorothiazide, à l'hydroxyfentanyl, à l'irbesartan, à la lidocaine, à la lidocaine n-oxide, à la morphine, au nebivolol, au nordiazépam, à la norfentanyl, à la norlidocaine, à l'ohmefentanyl, à l'olanzapine, à l'omega-hydroxynorfentanyl, à l'oméprazole sulphone, à l'oxazépam, au paracétamol, au propranolol, au tetrazépam, au valsartan et au zolpidem ont été supprimés, car considérés comme des contaminants d'origine pharmacologique.

3.4. Analyse en composantes principales (non supervisée)

Les résultats de l'analyse en composantes principales sont présentés dans la Fig. 2. L'analyse a révélé qu'il existait une variabilité lot à lot dans nos données, avec une séparation entre deux groupes de 20 et 50 échantillons. Ces deux groupes ont fait l'objet de deux analyses CL-SM séparées, ce qui a pu entraîner les variations observées. Il n'a pas été mis en évidence d'échantillon présentant des caractéristiques aberrantes.

3.5. Prédiction de la captation du FDG par méthodes de classification supervisée

Les performances de l'analyse discriminante des moindres carrés partiels sont présentées dans la Fig. 3. Les meilleures performances étaient obtenues avec des modèles à 3 composantes explicatives. La Fig. 4 représente la projection des échantillons en fonction des valeurs des deux premières composantes des modèles.

Les 20 métabolites présentant les poids les plus importants pour la première composante de chaque modèle sont présentés dans la Fig. 5. La N1,N12-diacetylspermine et la proline avaient les poids les plus importants en fonction du type de réduction utilisée.

3.6. Analyses comparatives univariées

Les analyses univariées réalisées ont montré que la N1,N12-diacetylspermine présente une forte association statistique ($p = 1,17E-7$) et un fort pouvoir prédictif pour l'intensité de fixation du FDG (aire sous la courbe sensibilité/spécificité : 0,83). La N1-acetylspermine, la forme mono-acétylée, présente également une forte association statistique ($p = 1,7E-5$). Ces résultats n'ont pas été vérifiés sur des données de métabolomique concernant 30 patientes niçoises prises en charge pour des cancers du sein et ayant eu un examen TEP-TDM au ^{18}FDG avant

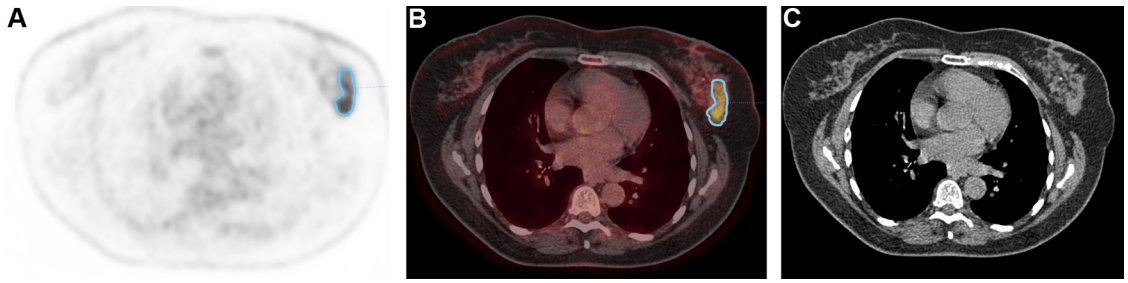

Fig. 1. Exemple de mesure de la SUV. Tumeurs des quadrants externes du sein gauche. Les contours de la région d'intérêt sont délimités en bleu. A. Image de tomographie par émission de positons. B. Image fusionnée de la tomographie par émission de positons et de la tomodensitométrie. C. Image de tomodensitométrie.
An example of SUV_{max} measurement. This example concerns a tumor of the upper quadrants of the left breast. The region of interest is delineated in blue. A. Fluorodeoxyglucose positron emission tomography image. B. Hybrid PET-CT image. C. Computed tomography image.

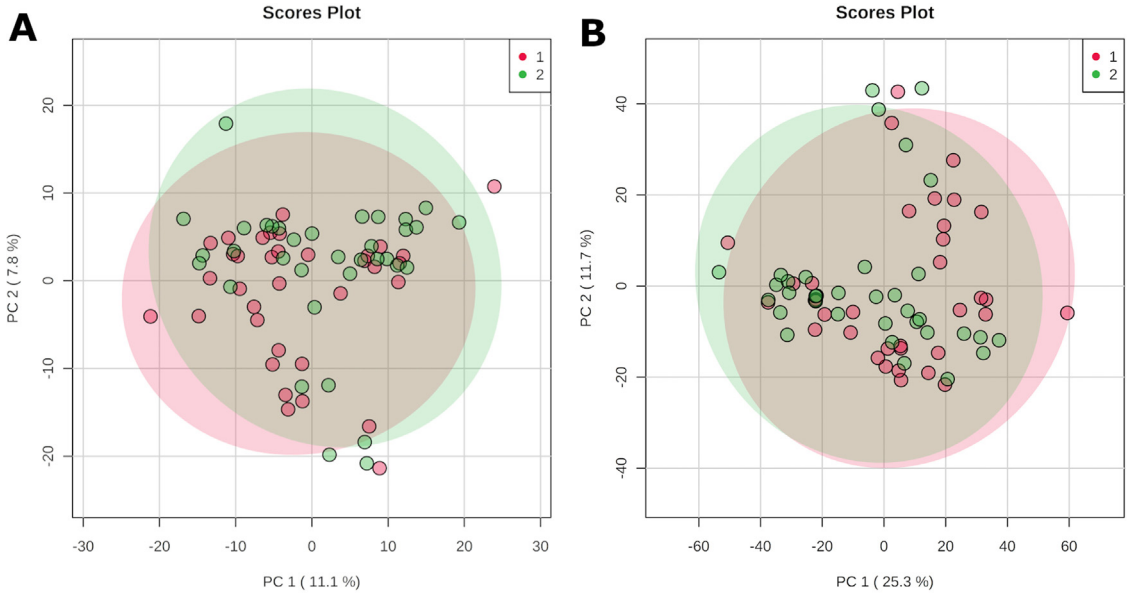

Fig. 2. Résultats des analyses en composantes principales. Projection des échantillons selon les axes des 2 premières composantes principales. Les zones rouges et vertes représentent les intervalles de confiance à 95 % des groupes « peu avides en FDG » et « très avides en FDG » respectivement. A. Données centrées. B. Données centrées et réduites avec la méthode Autoscaling. Avec cette analyse, les patientes sont regroupées sans prendre en compte le groupe d'avidité en FDG. Seules les données de métabolomique sont prises en compte. Nous pouvons voir que les patientes ne sont pas spontanément regroupées en groupes d'avidité en FDG différente. Il semble exister deux groupes, indépendants de l'avidité en FDG. Ceci peut être expliquer par une variabilité lot à lot (avec 2 lots de patientes). Les deux types de réductions sont à l'origine de résultats similaires.
Principal component analysis results. Projection of the samples on the first 2 principal components. The red and green regions represent the 95% confidence intervals for the "low FDG uptake" and "high FDG uptake" groups of patients. A. Mean centered and autoscaled data. B. Mean centered and Pareto-scaled data. In this analysis, samples are grouped into clusters without taking into account the FDG uptake. We can see that patients are not spontaneously clustered into groups with different FDG uptakes. Patients seem to be separated into two groups, independently of FDG uptake, which can be due to a batch effect in this case. The two different scaling techniques produce the same results.

traitement. On ne notait que des taux non significativement plus élevés de N1,N12-diacetylspermine pour les tumeurs « avides en FDG » (p = 0,21).

Plusieurs métabolites issus du métabolisme du tryptophane présentent également une association statistique avec l'intensité de fixation du FDG. En effet, les taux de L-kynurenine (p = 1,7E-5), d'acide kynurenique (p = 1,1E-4) et de L-formyl-kynurenine (p = 4E-3) sont plus élevés dans les tumeurs « avides en FDG » que dans les tumeurs « peu avides ». Ceci a également été constaté sous forme de tendance concernant la L-kynurenine dans la seconde cohorte (p = 0,15).

Par ailleurs, certains acides aminés comme le glutamate et la proline sont significativement plus élevés dans les tumeurs « avides en FDG » (p = 0,002 et 0,003 respectivement). De même, la glutathione, un dérivé du glutamate impliqué dans la réponse au stress oxydatif, était non significativement augmentée dans les tumeurs « avides en FDG » (glutathione oxydée : p = 0,06). Ceci a également été constaté dans la seconde cohorte (p = 0,02 et 0,04 respectivement).

Enfin, l'acetyl-carnitine et certains acides aminés acétylés comme la N-acetyl-alanine sont significativement plus élevés dans les tumeurs « avides en FDG » (p = 0,02 et 0,006 respectivement). Ceci a également été constaté dans la seconde cohorte (p = 0,015).

Les analyses univariées n'ont pas montré d'association statistique significative entre l'avidité en FDG et les métabolites de la voie de la glycolyse.

4. Discussion

4.1. Métabolisme énergétique

La glycolyse n'a pas pu être directement explorée, car un nombre insuffisant de métabolites de cette voie a été identifié dans nos données. Ceci peut être dû à la fragilité relative des petites molécules polaires que sont les sucres comparativement aux molécules plus robustes que sont, par exemple, les acides aminés. Les tumeurs étudiées n'ayant pas été analysées immédiatement après prélèvement, les sucres ont pu être dégradés.

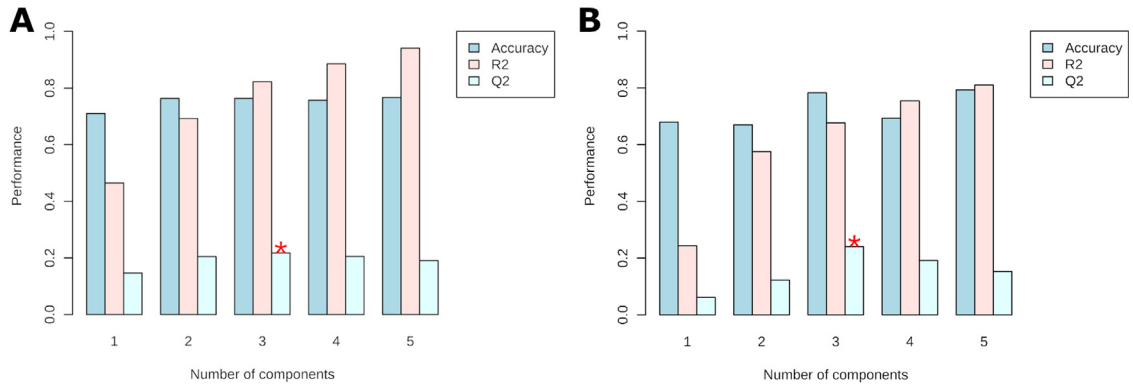

Fig. 3. Performances des analyses discriminantes des moindres carrés (PLSDA). Représentation des performances des modèles issus des analyses PLSDA pour différencier les groupes « peu avide en FDG » et « très avides en FDG » en fonction du nombre de composants utilisés. Les barres bleues, roses et bleu clair représentent la précision, le R2 et le Q2 respectivement. Les étoiles rouges indiquent le Q2 le plus élevé. A. Données centrées. B. Données centrées et réduites avec la méthode Autoscaling. Avec cette analyse le but est de créer un modèle regroupant les patientes en groupes d'avidité en FDG à partir des données de métabolomique. Le R2 est le coefficient de détermination du modèle. Il indique à quel point le modèle explique correctement la séparation des groupes. La précision correspond au pourcentage de patientes bien classées après validation croisée. Le Q2 indique à quel point la classification est robuste. Il peut varier de −1 à 1. Dans notre cas, le Q2 est positif mais faible. Ceci peut être du au faible nombre de patientes incluses.
Performance of Partial Least Square Discriminant Analysis (PLSDA) based models. Representation of accuracy (dark blue), R2 (pink) and Q2 (light blue) depending of the number of components included in the model. Red stars indicate the highest Q2. A. Mean centered and Autoscaled data. B. Mean centered and Pareto-scaled data. In this analysis, samples are grouped into clusters taking into account the FDG uptake. The R2 represents the fit of the model. Here, the model is well fitted with the Autoscaled data while the fitting is weaker with the Pareto-scaled data. The accuracy represents the percentage of samples that are correctly classified after cross validation. Here the accuracy varies between 0.75 and 0.8. The Q2 represents the confidence with which the samples are classified, it can vary from −1 to 1. Here the Q2 is positive but weak. This low value can be due to the relatively low number of samples and the fact that the FDG uptake is reduced to two groups.

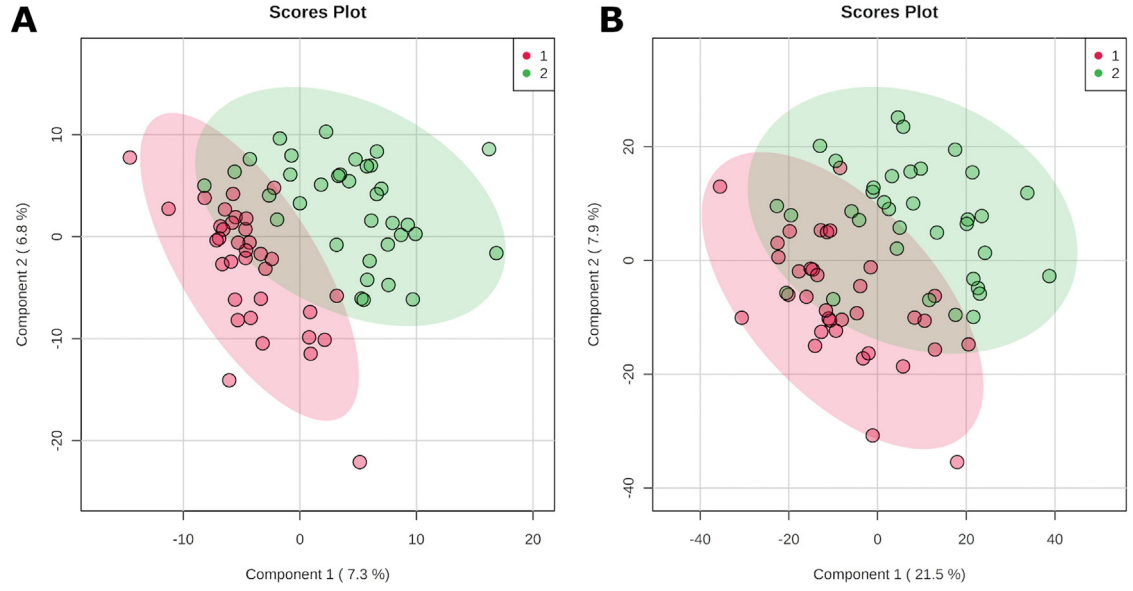

Fig. 4. Projection des échantillons selon les axes des deux principales composantes. Les points rouges correspondent aux tumeurs du groupe « peu avides en FDG », les points verts correspondent aux tumeurs du groupe « très avides en FDG ». A. Données centrées et réduites avec la méthode Autoscaling. B. Données centrées et réduites avec la méthode Pareto. La qualité de la séparation des groupes est liée au R2. Dans notre cas, il existe une séparation imparfaite des groupes. Celle-ci est meilleure avec les données réduites par Autoscaling qu'avec la méthode de Pareto.
PLSDA projections. Projection of the samples on the first 2 principal components. A. Mean centered and Autoscaled data. B. Mean centered and Pareto-scaled data. The separation of the data is linked to the model. If the model is well fitted, the data will be well separated. We can see that an imperfect separation is made between the two groups. The separation is better with the autoscaled data compared to the Pareto scaled data.

On peut également supposer que l'absence de résultats significatifs pour cette voie soit liée à l'absence de goulot d'étranglement au cours du métabolisme glucidique. Les métabolites pour lesquels des différences significatives ont pu être mises en évidence peuvent représenter les produits de réactions enzymatiques ou transports sur- ou sous-exprimés en fonction de l'activité tumorale, mais pour lesquels la suite de la cascade enzymatique est relativement moins exprimée, entraînant une accumulation de produit à une étape de la chaîne. Si l'ensemble de la voie de la glycolyse aérobie est suractivé et que le lactate ainsi formé est éliminé du tissu tumoral, les quantités de métabolites au sein de la tumeur peuvent être faibles, ce qui expliquerai nos résultats.

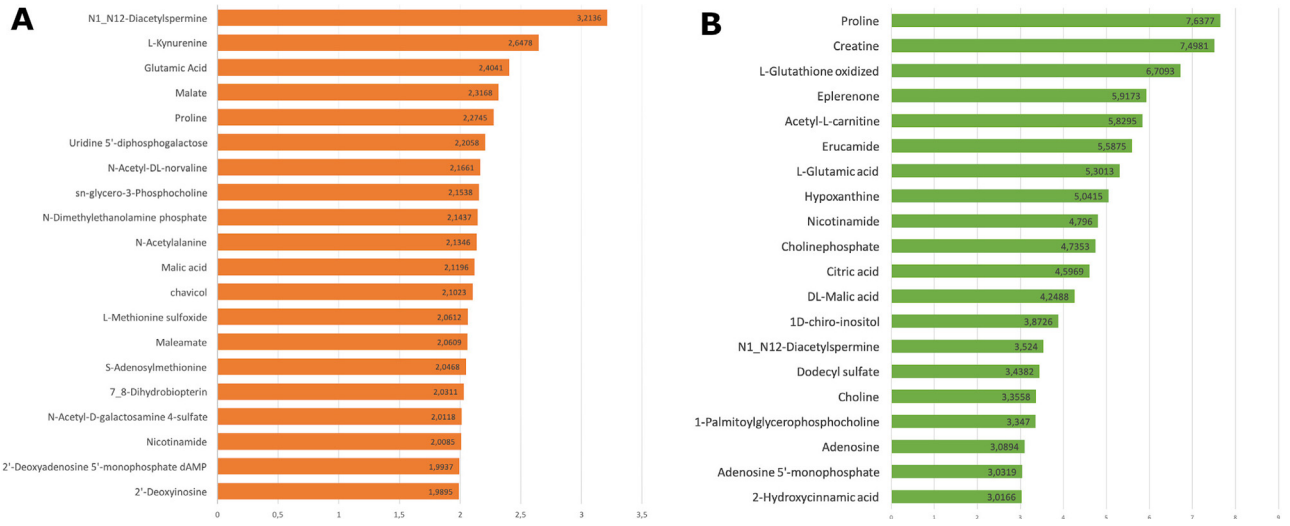

Fig. 5. Vingt métabolites ayant les plus grands poids dans les modèles issus des analyses PLSDA. A. Données centrées et réduites avec la méthode Autoscaling. B. Données centrées et réduites avec la méthode Pareto. La réduction par Autoscaling permet aux métabolites de faibles intensités d'avoir un poids équivalent aux métabolites de fortes intensités. Ici par exemple, la N1,N12-diacetyl-spermine, un métabolite de faible intensité, a un poids élevé. La méthode de Paretto ne réduit que partiellement la différence entre les métabolites de faibles et de fortes intensités. Ainsi les métabolites de fortes intensités auront tendance à avoir un poids plus élevé. Dans notre cas, avec la méthode de Paretto, bien que la N1,N12-diacetyl-spermine garde un poids élevé, la proline, un métabolite de forte intensité, a le poids le plus élevé.
Twenty most important metabolites for each PLSDA based models. Values represent the weights of each metabolite in the first component of each model. A. Mean centered and autoscaled data. B. Mean centered and Pareto-scaled data. Autoscaling enables features of small intensities to have equivalent weights compared to features of high intensities. Here for example a feature of low intensity, N1,N12-diacetyl-spermine has the highest weight. Pareto-scaling partially reduces the difference between high intensity features and small intensity features, therefore high intensity features still tend to have higher weights than low intensity features. Here for example, while N1,N12-diacetyl-spermine still has a high weight, Proline, a high intensity feature has the highest weight.

4.2. Voies métaboliques d'intérêt

Plusieurs métabolites d'intérêt ont pu être mis en évidence par notre approche. Ceux-ci présentent des liens plus ou moins étroits avec la glycolyse, résumés dans la Fig. 6.

La glutathione est un tripeptide dérivant du glutamate, de la glycine et de la cystéine jouant un rôle clé dans la réponse au stress oxydatif. La glutathione sous forme oxydée peut participer à la prédiction de l'avidité en FDG et est plus élevée dans les tumeurs « avides en FDG ». Ces résultats sont concordants avec des résultats rapportés concernant des niveaux élevés de glutathione dans des tumeurs du sein comparativement aux tissus sains [13]. Une augmentation du rapport glutathione oxydé/glutathione réduite a également été rapportée dans les tumeurs du sein, ce qui peut expliquer l'importance de la glutathione oxydée dans notre modèle [14]. Un haut niveau de glutathione oxydée peut refléter une production importante de dérivés réactifs de l'oxygène qui contribue à la prolifération tumorale, à une modification du métabolisme tumoral et des signaux cellulaires ainsi qu'aux mécanismes de résistance [15]. Des résultats discordants ont été rapportés concernant la corrélation entre les taux de glutathione et des facteurs pronostiques. Nos résultats suggèrent plutôt une corrélation positive entre agressivité tumorale et taux de glutathione oxydée.

Il existe une relation connue entre le stress oxydatif et le catabolisme des polyamines dont les produits finaux sont la N1,N12-diacetylspermine et la N1-acetylspermine [16]. Celles-ci peuvent participer à la prédiction de l'avidité en FDG et sont plus élevées dans les tumeurs « avides en FDG ». Des analyses complémentaires de nos données ont pu montrer que le taux de N1,N12-diacetylspermine est associé à d'autres critères pronostiques comme le grade SBR ($p = 0{,}018$), le taux d'atypies cellulaires ($p = 0{,}018$) et le taux de mitoses ($p = 0{,}083$).

La N1-acetylspermine et la N1,N12-diacetylspermine sont des produits de l'acétylation de la spermine et de la N1-acetylspermine par la Spermine/Spermidine Acetyl Transferase (SSAT), une enzyme limitante du catabolisme des polyamines. Elles sont également des substrats de la Polyamine Oxidase (PAO), une seconde enzyme du catabolisme des polyamines. Des taux plus élevés de SSAT et moins élevés de PAO ont été rapportés dans des tumeurs du sein comparativement au tissu mammaire non tumoral. Une association a également été mise en évidence avec des facteurs pronostiques comme le grade histologique ou la taille tumorale [17]. Ainsi, si l'intensité de captation tumorale du FDG est considérée comme un marqueur d'agressivité tumorale, ces résultats sont cohérents avec les observations précédemment rapportées. Paradoxalement, il a été rapporté que l'activation de SSAT et l'inhibition de PAO induisent une diminution de la prolifération cellulaire au niveau de cellules de cancers colorectaux et d'hépatocarcinomes [18]. Le rôle des polyamines dans la tumorogenèse reste incomplètement élucidé.

Le glutamate et la proline peuvent participer à la prédiction de l'avidité en FDG et sont plus élevés dans les tumeurs « avides en FDG », ceci peut être dû à une activité plus importante de l'axe glutamine-proline. Cette augmentation d'activité a déjà été rapportée dans des tumeurs du sein plus agressives comparativement à des tumeurs moins agressives [19] et peut être due à des besoins plus importants en acides aminés protéinogènes dans les tumeurs à prolifération rapide. Probablement pour les mêmes raisons, la tyrosine et la N-acetyl-methionine sont plus élevées dans les tumeurs « avides en FDG » que dans les tumeurs « peu avides en FDG ».

La L-acetyl-carnitine peut participer à la prédiction de l'avidité en FDG et est plus élevée dans les tumeurs « avides en FDG » que dans les tumeurs « peu avides en FDG ». Ceci peut être expliqué par une hyperactivation du cycle de la carnitine en réponse à l'acidification du milieu due à la glycolyse aérobie et à la production de lactate [20].

Des métabolites issus du métabolisme du tryptophane, notamment la L-kynurenine, peuvent participer à la prédiction

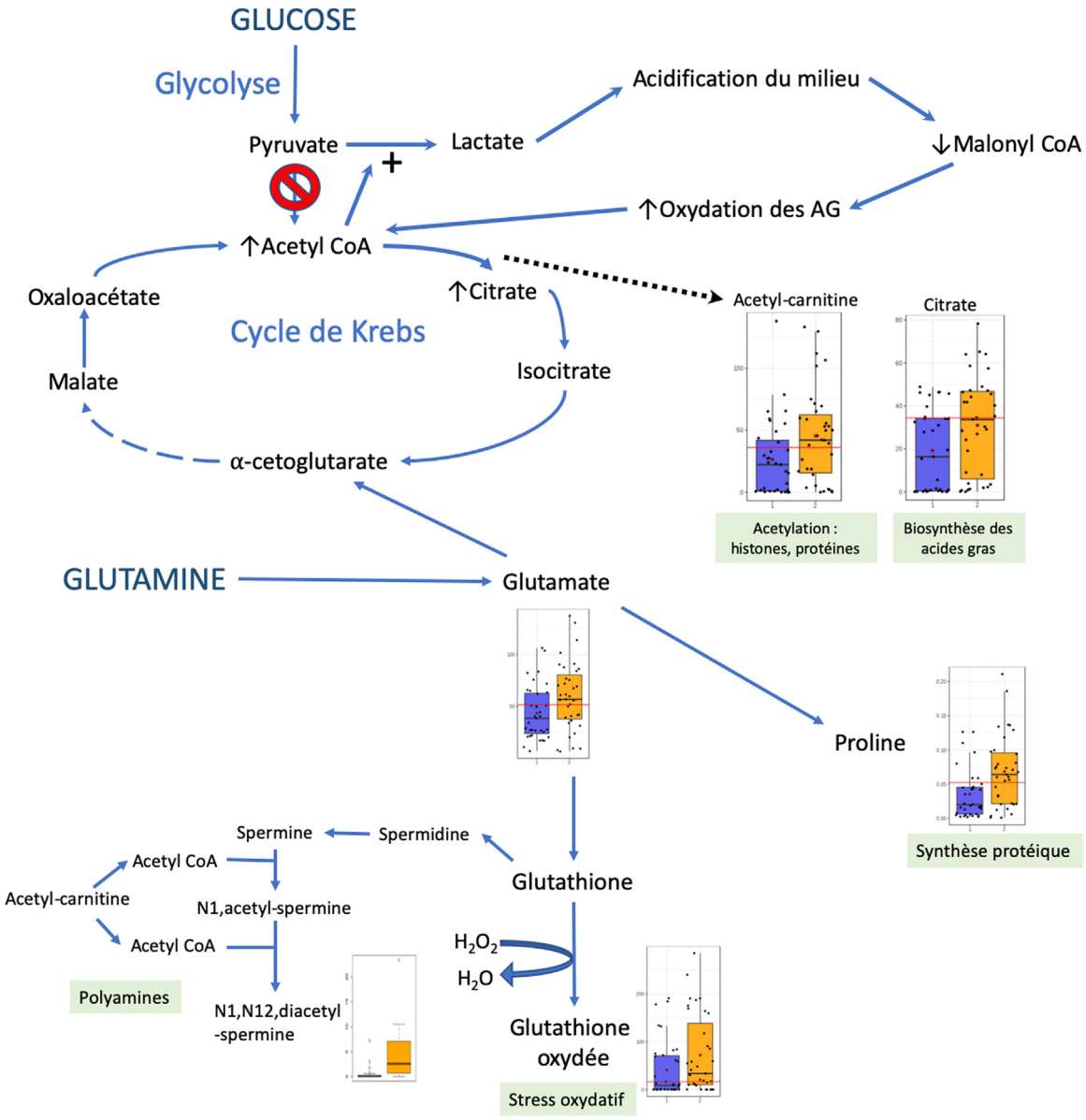

Fig. 6. Schémas simplifiés des différentes voies métaboliques mentionnées. Les valeurs des métabolites d'intérêt sont représentées sous forme de boîtes à moustaches. Les boites bleus correspondent au groupe des tumeurs à faible avidité en FDG et les boites oranges correspondent au groupe des tumeurs à forte avidité en FDG. Les métabolites sont représentés au sein des voies métaboliques ayant des altérations en lien avec l'avidité en FDG.
A simplified view of the discussed metabolic pathways. Metabolites of interest are represented using box-plots. Blue boxes represent low FDG uptake tumors and orange boxes represent high FDG uptake tumors. Metabolites are represented within the energetic pathways considered to be altered depending on FDG uptake.

de l'avidité en FDG et sont plus élevés dans les tumeurs « avides en FDG ». L'hyperactivation de l'axe de la kynurenine a été rapportée comme un mécanisme d'échappement tumoral au système immunitaire et il a été montré que cet axe était plus actif dans les tumeurs du sein plus agressives [21]. Des traitements visant à inhiber la première enzyme de l'axe de la kynurenine, IDO-1 sont actuellement en phase de test. Si ces traitements s'avèrent efficaces, une SUV_{max} élevée pourrait être un facteur prédictif de sensibilité tumorale.

4.3. Limites de l'étude

4.3.1. Biais lié à une variabilité lot à lot

Les ACP ont montré la présence d'une variabilité lot à lot dans nos données. Cette variabilité pourrait être due au changement de colonne de chromatographie entre les passages de ces deux lots d'échantillons. L'utilisation de standards internes au cours des analyses CL-SM aurait pu limiter cette variabilité. Celle-ci peut représenter un biais de confusion dans nos résultats. Certains résultats ont pu être validés en confrontant nos résultats à une seconde base de données et ne sont donc probablement pas le résultat de cette variabilité. Cependant, cette seconde base étant de plus petite taille, il n'a pas été possible de valider tous nos résultats. On peut espérer que le fait que les deux classes étudiées aient été présentes dans des proportions équivalentes dans les deux lots présents limite l'influence de cette variabilité lot à lot.

4.3.2. Données manquantes concernant les apports exogènes

Les données de métabolomique traduisent l'influence de facteurs endogènes et exogènes sur les processus cellulaires. Les

données concernant le régime alimentaire ou les prises médicamenteuses des patientes étudiées n'étaient pas disponibles pour cette étude. Il aurait été préférable de pouvoir confronter les données de métabolomique à ces informations. La tangeritine présente, par exemple, une association statistique avec l'avidité en FDG. Il s'agit d'un métabolite présent dans de nombreux fruits et légumes et pour lequel des propriétés cytostatiques ont été rapportées dans le cadre du cancer du sein [22]. Elle est désormais proposée sous forme de complément alimentaire. Paradoxalement, les taux étaient plus élevés dans les tumeurs « très avides en FDG ». Ce résultat est difficilement interprétable en l'absence d'information concernant les apports exogènes des patientes en tangeritine.

4.3.3. Complexité des données

Le métabolome représente un vaste réseau de molécules interconnectées par le biais d'enzymes. Au cours de ce travail, nous nous sommes intéressés à une portion de ce vaste réseau. Bien qu'il existe désormais des outils informatiques permettant de mieux appréhender ces données, leur interprétation reste complexe, notamment, car certaines petites molécules et enzymes ne sont pas encore connues. De plus, chaque petite molécule peut à la fois être produit et substrat de plusieurs enzymes dans ce réseau et il peut être difficile d'interpréter les mécanismes à l'origine des variations observées. Des analyses supplémentaires de protéomique pourraient permettre de confirmer ou infirmer les hypothèses émises à partir des données de métabolomique.

5. Conclusions et perspectives

La métabolomique offre accès à un grand nombre de données intégrant l'influence de l'ensemble des facteurs internes et externes modifiant les processus cellulaires. La compréhension des phénomènes complexes influençant ces données nécessite des transformations de données et des approches statistiques innovantes.

Notre approche n'a pas permis de mettre en évidence de corrélations entre la glycolyse et l'avidité tumorale en FDG. Ceci pourrait être dû à la dégradation précoce de ces métabolites fragiles dans le tissu tumoral.

Cependant, il existe des corrélations entre d'autres métabolites et l'avidité tumorale en FDG dans le cadre du cancer du sein. La valeur pronostique de certains de ces métabolites a déjà été rapportée dans la littérature. Certains métabolites corrélés à l'avidité tumorale en FDG sont pour l'instant inconnus ou peu étudiés. L'étude de ces métabolites pourrait permettre d'identifier de nouveaux facteurs pronostiques dans le cadre du cancer du sein.

Par ailleurs, l'utilisation de la mesure in vivo de l'avidité tumorale en FDG, comme outil de classification des échantillons dans le cadre des études de métabolomique visant à rechercher des biomarqueurs, offre une alternative à la comparaison « tissus sains », « tissus malades » permettant d'éviter des prélèvements invasifs chez des patients indemnes. Cette approche pourrait être employée pour d'autres types de cancer.

Déclaration de liens d'intérêts

Les auteurs déclarent ne pas avoir de liens d'intérêts.

Annexe A. Matériel complémentaire

Le matériel complémentaire accompagnant la version en ligne de cet article est disponible sur : http://doi.org/.10.1016/j.mednuc.2020.03.002.

Références

[1] De Nicola GM, Cantley LC. Cancer's fuel choice: new flavors for a picky eater. Mol Cell 2015;60:514–23.
[2] Warburg O, Wind F, Negelein E. The metabolism of tumors in the body. J Gen Physiol 1927;8:519–30.
[3] Heiden MGV, Cantley LC, Thompson CB. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science 2009;324:1029–33.
[4] Liberti MV, Locasale JW. The Warburg effect: how does it benefit cancer cells? Trends Biochem Sci 2016;41:211–8.
[5] Brown RS, Leung JY, Fisher SJ, et al. Intratumoral distribution of tritiated-FDG in breast carcinoma: correlation between Glut-1 expression and FDG uptake. J Nucl Med 1996;37:1042–7.
[6] Groheux D, Hindié E, Salaün PY. Cancers du sein. Med Nucl 2019;43:85–103.
[7] Gallagher BM, Fowler JS, Gutterson NI, et al. Metabolic trapping as a principle of oradiopharmaceutical design: some factors responsible for the biodistribution of [18F] 2-deoxy-2-fluoro-D-glucose. J Nucl Med 1978;19:1154–61.
[8] Diao W, Tian F, Jia Z. The prognostic value of SUVmax measuring on primary lesion and ALN by 18F-FDG PET or PET/CT in patients with breast cancer. Eur J Radiol 2018;105:1–7.
[9] García Vicente AM, Soriano Castrejón Á, León Martín A, et al. Molecular subtypes of breast cancer: metabolic correlation with 18F-FDG PET/CT. Eur J Nucl Med Mol Imaging 2013;40:1304–11.
[10] Courant F, Antignac J-P, Dervilly-Pinel G, et al. Basics of mass spectrometry based metabolomics. Proteomics 2014;14:2369–88.
[11] McCartney A, Vignoli A, Biganzoli L, et al. Metabolomics in breast cancer: a decade in review. Cancer Treat Rev 2018;67:88–96.
[12] Berg RA, van den, Hoefsloot HC, Westerhuis JA, et al. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 2006;7:142.
[13] Gamcsik MP, Kasibhatla MS, Teeter SD, et al. Glutathione levels in human tumors. Biomarkers 2012;17:671–91.
[14] Perquin M, Oster T, Maul A, et al. The glutathione-related detoxification system is increased in human breast cancer in correlation with clinical and histopathological features. J Cancer Res Clin Oncol 2001;127:368–74.
[15] Galadari S, Rahman A, Pallichankandy S, et al. Reactive oxygen species and cancer paradox: to promote or to suppress? Free Rad Biol Med 2017;104:144–64.
[16] Casero RA, Stewart TM, Pegg AE. Polyamine metabolism and cancer: treatments, challenges and opportunities. Nat Rev Cancer 2018;18:681–95.
[17] Wallace HM, Duthie J, Evans DM, et al. Alterations in polyamine catabolic enzymes in human breast cancer tissue. Clin Cancer Res 2000;6:3657–61.
[18] Wang C, Ruan P, Zhao Y, et al. Spermidine/spermine N1-acetyltransferase regulates cell growth and metastasis via AKT/β-catenin signaling pathways in hepatocellular and colorectal carcinoma cells. Oncotarget 2016;8:1092–109.
[19] Craze ML, Cheung H, Jewa N, et al. MYC regulation of glutamine–proline regulatory axis is key in luminal B breast cancer. Br J Cancer 2018;118:258–65.
[20] Melone MAB, Valentino A, Margarucci S, et al. The carnitine system and cancer metabolic plasticity. Cell Death Dis 2018;9:228.
[21] Heng B, Lim CK, Lovejoy DB, et al. Understanding the role of the kynurenine pathway in human breast cancer immunobiology. Oncotarget 2015;7:6506–20.
[22] Morley KL, Ferguson PJ, Koropatnick J. Tangeretin and nobiletin induce G1 cell cycle arrest but not apoptosis in human breast and colon cancer cells. Cancer Lett 2007;251:168–78.